INNOVATIONS IN
The biggest questions in science

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



Butterfly effect

Penchant for exotic plant pushed local population to extinction **PAGES 171 & 238**

MENTAL HEALTH

DEALING WITH DEPRESSION

Stories of hope from the darker side of research

PAGES 160 & 267

PARTICLE PHYSICS

POSITIVE RESULT

Precision measurement of the proton's weak charge

PAGES 171 & 207

NEUROSCIENCE

DEFENSIVE ACTION

The brain circuits that control how mice respond to threats

PAGES 172 & 183

NATURE.COM/NATURE

10 May 2018

Vol. 557, No. 7704

THIS WEEK

EDITORIALS

MENTAL HEALTH Awareness is good but campus action is better **p.140**

WORLD VIEW We must label harassment as academic misconduct **p.141**



PERIWINKLE Enzymes that make cancer drug identified in flower **p.143**

Harassment on China's campuses

A string of cases at Chinese universities shows a system that is trying to address concerns and implement change, but still has some way to go.

Some 20 years ago, a student at Peking University in Beijing took her own life after making allegations that she had been sexually harassed and raped by a professor. Her case made little impact at the time, but it is doing so now. The issues it raises highlight two points about how the #MeToo movement is now playing out on the campuses of Chinese universities. It shows the extent to which things have changed, and underlines ways in which these changes do not yet go far enough.

The student was called Gao Yan. When friends and supporters last month highlighted the anniversary of her death, Peking University confirmed that an investigation at the time had criticized Shen Yang, a famed linguist at the university, for having an inappropriate six-month relationship with her, which he ended nine months before her suicide. Shen could not be reached by *Nature*. He told Chinese media the accusations against him were untrue. After the case was made public last month, he was fired from his post at Shanghai Normal University.

Peking University, which Shen left in 2011, also posted previously unreleased documents disclosing some details of its 1998 investigation into the case. And it published statements that noted recent efforts, including the introduction in 2016 of a 'teacher's handbook', to reinforce the ethics of its professors and draft regulations concerning sexual harassment on campus (currently under consideration). The institution deserves some credit for these efforts, albeit two decades on. There are many other universities — in China and elsewhere — that stick their heads deep in the sand when controversy arises, on sexual harassment, scientific misconduct or other matters.

That the university felt the need to respond to public pressure at all, never mind to issue statements and attempts at reassurance over a historic case, gives some indication of how things are changing for the better in China. Awareness of harassment, and intolerance for harassers, is certainly on the rise there, as in many places. A string of well-publicized sexual-harassment cases has hit university campuses in China in recent months, and senior academics accused of improper behaviour have lost their positions or faced other sanctions.

The shift goes beyond academia, too. Some technology companies have been forced to apologize for discriminatory hiring policies that target attractive women, and for advertisements that boast about the beauty of their female employees.

Together, the public airing of these cases is a positive development. Unsavoury activity is being exposed and, to some extent, those found guilty are receiving penalties that might deter others from similar behaviour. This could signal that China is reaching a new stage of transparency, where such issues can be discussed and sexual harassment will no longer be tolerated. But there are still many reasons to be concerned.

Following the response from Peking University over the Gao Yan case, a group of current students there pressed the university for more details. Since then, one of them — Yue Xin — has complained on social media that university officials have pressured her to stop asking for the information. Her allegations have made headlines and yielded statements of support from around the world. They have partly overshadowed official celebrations of the university's 120-year anniversary this month. And, in response, the university's newfound openness is faltering — numerous attempts to contact the institution have gone unanswered. Meanwhile, students say that posters they have put up around campus voicing support for Yue have been quickly removed.

"Awareness of harassment, and intolerance for harassers, is certainly on the rise."

Chinese President Xi Jinping visited Peking University last week. Ironically, he praised it as the birthplace of the 1919 May Fourth movement, a series of student protests that triggered wide social and political unrest and ultimately produced the nation's communist leaders.

Concerns over the response to the student protests must be addressed. It is one thing for universities to state that sexual harassment will not be tolerated. It is another entirely for them to buy into the kind of wholesale changes in regulations, behaviour and attitude that are required. The #MeToo movement is growing into an irresistible force. Now is not the time for universities — in China or elsewhere — to act like immovable objects. ■

Crowd scene

Satisfying symmetry between an unusual test of quantum physics and peer review of its findings.

Wednesday, 30 November 2016, saw crowds of Cubans line the streets of Havana to say a final farewell to Fidel Castro. On the same day, fans gathered to watch Spanish football minnows Hércules hold giants Barcelona to a surprise one-all draw in their Copa del Rey tie. And a crowd of 100,000 people around the world came together over an online video game to type in a series of 0s and 1s as fast as their fingers could fly, to put to the test a central feature of quantum mechanics.

The gamers were part of a day-long experiment called the BIG Bell Test, the results of which are described on page 212 (The BIG Bell Test Collaboration. *Nature* **557**, 212–216; 2018). The findings were produced in an unusual way and, with the symmetry that one would expect from a beautiful physics theory, they were reviewed for publication in an unusual way, too.

Less than a year after the test, the BIG Bell Test paper landed for review from *Nature* on the desk of Sabrina Maniscalco, a physicist at the University of Turku in Finland. Seeing the crowdsourced public input it contained, Maniscalco decided to crowdsource the requested

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

review — well, up to a point. She showed it to seven of her senior PhD students, and then invited them to discuss it with her over a day of brainstorming, literature search and pizza. She channelled the resulting comments and criticisms into the referee's report that she submitted, and which all seven students saw again before submission. No confidence was breached during the review process: the students agreed to strict confidentiality, and Maniscalco opted in to our referee-accreditation trial scheme, which means that her name, together with those of the participating students, appears in a statement at the end of the research paper. Two other reviewers offered their thoughts on the paper in the usual way.

It is not unusual for a lab head to delegate a paper's peer review to a junior colleague, or to collaborate with one — and, unfortunately, the colleague's contribution isn't always acknowledged. It is unusual, as far as we know, to make peer review a team exercise and to openly state that that was how it was done. It's not an experiment we expect will be repeated often. Still, contrary to popular belief, our editors are, in principle, happy for referees to involve others in the review process, as long as confidentiality is assured and the editors are kept in the loop. It's right for all those involved in peer review to be acknowledged. At its best, such collaboration can enrich the review process and help junior researchers develop the skills needed to become effective referees themselves.

In this case, the paper explores the tension between quantum physics and local realism. The latter brings together two principles: locality — according to which, observing a particle at one physical location cannot have immediate effects on the properties of a particle at a different location — and realism, which expresses how the observable features of particles exist even if we don't actively measure them. But in quantum mechanics, correlations between distant particles exist that are so strong they violate local realism. Put differently, in quantum theory it is possible to have two correlated particles far away from each other, to measure the first and, as a result, learn something about the second without having observed it directly.

So here's the conundrum: does quantum mechanics really violate local realism, or could it be the case that some unknown factors would complete the theory and explain these apparent violations? In the 1960s, the physicist John Bell offered a way to tackle the problem in the laboratory, by studying quantum correlations in the form of entanglement.

"The BIG Bell Test closes the 'freedom-of-choice' loophole."

In these experiments, sequences of spatially separated measurements on entangled particles lead to computing a quantity that can have values not possible in the context of local and realistic theories. Bell tests have confirmed the validity of quantum theory many times, but they include assumptions that leave wiggle room for non-quantum explanations as to why local realism is violated, and so physicists have been looking for ways to close these loopholes ever since.

In 2015, physicists showed that successful Bell tests could not be due to speed-of-light communication between the particles, or to inefficient detection processes during the measurements. But another, more subtle, loophole was still open. Bell tests also assume that experimenters have free choice over which measurements they make on each particle. And yet, hidden parameters could be influencing these choices to produce correlations that give the illusion of entanglement.

The BIG Bell Test closes this freedom-of-choice loophole. The various experimental groups had no say in which measurement settings to use. Instead, they performed their measurements according to the unpredictable streams of bits received from the 100,000 gamers.

The results show the presence of correlations strong enough to contradict local realism. Maybe that's how 30 November 2016 might be remembered: the day the people of the world came together to test quantum theory. ■

Helping hands

Awareness of mental health must be matched with steps such as better training for supervisors.

It's Mental Health Awareness Month in the United States, and next week is Mental Health Awareness Week in the United Kingdom. And awareness is certainly on the rise. Already this year, reports, surveys and studies have highlighted psychological struggles experienced by the old, the young, schoolchildren, men, women, soldiers, immigrants and refugees, football players, dancers, actors, social-media users, musicians, and elite athletes immediately after the Olympic Games. As quoted often, one in four people have a mental-health condition. And rates of depression and anxiety reported by postgraduate students are unacceptably high. This week, *Nature* is working to improve awareness of how mental illness can affect researchers: in this issue's Careers (page 267) and Comment (page 160) sections, several scientists share their experiences with honesty and admirable courage.

Awareness on its own is obviously not enough. John Lennon wrote that life is what happens while you're busy making other plans. Well, a life with mental illness can feel a lot like something that happens while well-meaning people are busy raising awareness. So, how do we make sure that those affected actually feel heard, supported and better?

Nature is trying to play a small part. Last month, we received a staggering response from readers to a Careers item about the alarmingly high rates of mental-health concerns reported by postgraduates. We invited people to tell us their stories, which we collected through a confidential online form. Our editors hoped to find some examples of success that we could share. Yet, almost without exception, the outpouring of 300-plus stories we received were from people who wanted support but were getting little, if any. (Of course, those who have found support might be less likely to tell their tale.) We printed five respondents' stories, with their kind permission, last week (*Nature* **557**, 129–131; 2018).

Most who replied were postgraduate students and postdocs, but several established scientists also wrote in to point out that mental-health problems are not confined to the young. We want to thank all those who responded so openly: it was harrowing reading, and will help to drive our future coverage of these issues. As many people struggling with their mental health eventually realize, it rarely helps to keep quiet. Reach out to someone and you'll probably be surprised at how readily they acknowledge what you're going through. Perhaps they are, too.

One problem is that, according to a report last year by RAND Europe, "the evidence around the effectiveness of interventions to support the mental health of researchers specifically is thin. Few interventions are described in the literature and even fewer of those have been evaluated" (see go.nature.com/2juanaw).

Some efforts are already under way to help postgraduates, beyond simply raising awareness. In March, the Higher Education Funding Council for England announced it would put a total of £1.5 million (US$2 million) towards improving mental health at 17 universities. Several schemes will endeavour to better train and equip PhD supervisors to mentor their students. This is much needed: a dysfunctional supervisor–student relationship was a common complaint from many who wrote to us.

Manipal Academy of Higher Education in India has established an independent, confidential student-support centre so that students can come directly to ask professional psychologists for immediate help. Meanwhile, the Francis Crick Institute in London has more than 20 registered mental-health first aiders, who are trained to recognize mental-health issues, provide initial help and guide people towards professional services where appropriate. These and other examples of good practice and sources of support are collated in a dedicated page on our website (see go.nature.com/2i9a6yx). We hope they inspire more. ■

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.


ANN OLSSON

Harassment should count as scientific misconduct

Scientific integrity needs to apply to how researchers treat people, not just to how they handle data, says **Erika Marín-Spiotta**.

In the past year, allegations of egregious sexual harassment and even assault have emerged across the spectrum of science. *Nature* has already run several stories on the topic just this quarter.

When I talk to senior scientists, many view harassment as an injustice that happens somewhere else, not in their field or at their institution. But data suggest that the problem is ubiquitous. In separate surveys of tens of thousands of university students across Australia, the United Kingdom and the United States, upwards of 40% of respondents say that they have experienced sexual harassment. A survey last year by the US National Postdoctoral Association found that 28% of respondents reported experiencing at least one instance of harassment while they were trainees; offenders were predominantly reported as being faculty or staff members (go.nature.com/2ju83ox). Neither are faculty members safe from mistreatment by colleagues.

Research culture and policies are quick to denounce plagiarism, data fabrication and mismanagement of funds, yet we have too long ignored the mistreatment of people.

Science is a social endeavour; ignoring harassment perpetuates a culture in which people who experience or witness hostile behaviours are afraid to speak up, cannot do their best work, or leave science altogether. Last September, the American Geophysical Union (AGU) defined harassment, bullying and discrimination as scientific misconduct, and outlined consequences. The greater scientific community should do the same.

My colleagues and I are developing a programme to reduce harassment in the geosciences — one of the least diverse scientific disciplines — supported by the US National Science Foundation (NSF). The project, ADVANCEGeo, equips bystanders to respond to and prevent harassment in the field, lab, office and at conferences, and advocates for inclusion of the subject in courses on ethical research conduct.

Our efforts are part of a growing movement. Just this year, both the US NSF and the UK Wellcome Trust put forward policies targeting harassment by grant recipients. Many scientific societies have codes of conduct that specifically prohibit harassment at meetings, and some, such as the American Astronomical Society and the Geological Society of America, address the professional treatment of others in their general codes of ethics.

Harassment, bullying and discrimination damage science at the individual, community, institutional and societal levels. The behaviours cause health problems, fear, mistrust, depression and trauma. That results in decreased productivity and the exclusion of people who might have led highly satisfying scientific careers and made important contributions.

Less-represented populations are disproportionately affected. A 2017 study in astronomy and planetary science found that women of colour were more likely than other groups to report skipping professional events because they felt unsafe (K. B. H. Clancy *et al. J. Geophys. Res. Planets* **122**, 1610–1623; 2017). In a 2016 survey of physicists identifying as lesbian, gay, bisexual, transgender or queer (LGBTQ), one-third of respondents reported that they had considered leaving their institutions; this group was also more likely to have experienced or witnessed hostile behaviours (see go.nature.com/2wczfih). Other groups are also likely to be vulnerable, but data are sparse.

Why aren't the laws already in place against harassment sufficient? People cannot count on their enforcement, especially given that imbalances of power in academia favour the perpetrators. Fear of retaliation also keeps people from reporting to employers in the first place.

Defining harassment as misconduct provides more ways of deterring it. For example, the AGU has developed processes for investigating allegations, including for addressing concerns that might not rise to the level of a formal complaint. Sanctions might include being barred from meetings or publishing in society journals, or the denial of an award. To be clear, when talking about harassment, I am not referring to socially awkward interactions but to well-defined and documented behaviours — such as unwanted groping and requests for sexual favours — that create a hostile work environment.

IMPROVED ETHICS TRAINING SHOULD **LAY BARE** POWER DYNAMICS AND BEHAVIOURS THAT ALLOW **HARASSMENT.**

A better sense of how and how often harassment happens in science would help to convince the community of its pervasiveness, and counter mistaken beliefs that it is not common in the workplace. We also need more data on the experiences of people from under-represented groups, and on how hostile climates affect efforts to maintain diversity.

We must craft effective interventions. Courses in responsible research conduct, currently a requirement for trainees, should cover how to prevent and respond to harassment and bullying. Institutions should offer in-person training, including discipline-specific scenarios that people can relate to. Departments should communicate their workplace values publicly. This is starting to happen. For example, the geosciences department at Middle Tennessee State University in Murfreesboro has crafted a code of conduct that describes principles and practices for professional behaviour, adapted from the AGU's. Scientific societies have also made plans to offer bystander-intervention training at their conferences.

Improved ethics training across career levels should draw from the social and behavioural sciences to lay bare the power dynamics and behaviours that allow harassment, bullying, racism and sexism to persist in the sciences. The integrity of the scientific enterprise demands that we stop tolerating such behaviours. ■

Erika Marín-Spiotta *heads the Biogeography and Biogeochemistry Research Group at the University of Wisconsin–Madison, USA.*
e-mail: advancegeopartnership@gmail.com

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

SEVEN DAYS *The news in brief*

PEOPLE

Scientist detained

A British–Iranian professor is being detained in Iran following his arrest during a trip to attend an academic workshop in Tehran, according to Iranian media reports. Abbas Edalat, a computer scientist and mathematician at Imperial College London, was detained by Iran's Revolutionary Guards in April for his alleged role in spying activities. Reports suggest that the authorities confiscated his computer, notebooks and CDs. Edalat founded a pro-peace organization, the Campaign Against Sanctions and Military Intervention in Iran. Imperial College says it is "concerned for his welfare", and is urgently seeking further information about the incident.

PNAS resignation

On 1 May, Inder Verma, a cancer researcher at the Salk Institute for Biological Sciences in La Jolla, California, resigned as editor-in-chief of the journal *Proceedings of the National Academy of Sciences*. The move comes after the publication of an investigation by *Science*, in which several female researchers who were either at the institute or had ties to it between 1976 and 2016 allege that Verma harassed them. Verma, who served on powerful committees at the institute, vehemently denied the allegations in a statement to *Nature*. The Salk Institute suspended him on 21 April while it investigates the claims.

Biologist sacked

The University of Tokyo announced on 27 April that it has dismissed cell biologist Yoshinori Watanabe. The decision followed a university investigation that found manipulated images in five of Watanabe's papers, which the institution said amounted to scientific misconduct. One paper has been retracted, and two others have been corrected. During his career, Watanabe made substantial contributions to scientists' understanding of how cells divide. Watanabe says that the problems that were identified in the five papers did not affect the studies' conclusions. He says that he had no intention of misleading others, and believes he did not commit misconduct. Watanabe told *Nature* that he resigned on 28 February, before his official dismissal. He will now spend a year retraining in data acquisition and presentation at the laboratory of Nobel prizewinner Paul Nurse, at the Francis Crick Institute in London.



HINDUSTAN TIMES/GETTY

Deadly dust storms hit India

Severe dust storms occurring in the lead up to the monsoon season have killed more than 100 people in the Indian states of Uttar Pradesh and Rajasthan in the past week. Most deaths occurred when buildings collapsed in unusually intense winds that approached speeds of 130 kilometres an hour. On 3 May, India's Meteorological Department said that the storms followed 'intense heating' over the plains of northwestern India, with temperatures in the region reportedly reaching 45–50 °C. Meteorologists say that high temperatures might have intensified the storms. The high fatality rate was probably due to the storms striking while people slept.

EVENTS

Apology for ritual

The Chinese Academy of Sciences has apologized for a Taoist ceremony organized by construction workers that included the slaughter of a lamb to inaugurate the construction of a nuclear test reactor. Images showing a monk chanting while the lamb was killed sparked online controversy in China. The April ceremony marked the start of construction of a thorium molten salt reactor in north-central China, part of a 22-billion yuan (US$3.5-billion) project to be operated by the academy's Shanghai Institute of Applied Physics. In a statement posted on its homepage on 30 April, the academy said it apologized because the ritual was "against the spirit of science" and that institute staff did not know the ritual was planned.

US biobank begins

An ambitious project to track the health of one million people in the United States for at least ten years will start enrolling participants on 6 May. The US National Institutes of Health initiative — announced by former president Barack Obama during his 2015 State of the Union address — aims to recruit volunteers who range in geographical locale, ethnic background, health conditions and age. The biological information that researchers will collect from participants include blood samples, biometric data from devices such as Fitbits, and medical records. The project will

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

accept only people over age 18 initially, but plans to include children in the future.

Day Zero delayed

Cape Town, South Africa, has averted Day Zero — the date officials predicted the city's reservoirs would run too low to deliver potable water to the metropolis. Citizens have saved the city — for now — by reducing their water usage from 1.2 billion litres a day in late 2016 to 450 million litres a day this year, said Helen Zille, Cape Town's former mayor, on 3 May. She cautioned that the threat could return next year, and that its timing would depend on how much rain the region receives in August.

PUBLISHING

Nature's new editor

Geneticist Magdalena Skipper will be the next editor-in-chief of *Nature*, the journal announced on 2 May. Skipper (**pictured**), who is currently editor of the open-access journal *Nature Communications*, will be the first woman to head the 149-year-old title. She will take over from Philip Campbell, who has led *Nature* since 1995, on 1 July. Skipper says that during her tenure, she wants to continue the journal's work to ensure that scientific findings are reproducible and robust, particularly in the age of big data. She would also like *Nature* to focus more on early-career researchers. See go.nature.com/2hx8woa for more.

IAN ALDERMAN



ENVIRONMENT

Pollution report

Air pollution is decreasing in Chinese cities, but increasing in cities in Pakistan, Bangladesh and India. According to the latest global air-quality report by the World Health Organization (WHO), released on 2 May, cities in those three countries have experienced the steepest rises in air pollution since 2010. In China, the average annual levels of small, dangerous particles known as $PM_{2.5}$ (those with a diameter of 2.5 micrometres or less) have dropped by 17% since the 2017 report. This reduction was achieved by the government enforcing strict limits on industrial emissions, increasing renewable power and increasing access to cleaner fuels for indoor heating and cooking. But dozens of Chinese cities still rank among the 100 most polluted, and an estimated 2 million people die in China every year because of air pollution. The WHO survey reviewed air-quality statistics in 4,300 major cities for 2016. It found that more than 90% of the world's population breathes unsafe air.

CLIMATE

Carbon hits high

The average concentration of atmospheric carbon dioxide, a key greenhouse gas, for the month of April exceeded 410 parts per million, researchers announced on 2 May. This is the first time in recorded history that the monthly CO_2 average has breached this threshold. Climate scientists took the measurements at the Mauna Loa Observatory in Hawaii, where they have been monitoring atmospheric CO_2 levels since 1958. Average global temperature has already increased by 1 °C from the pre-industrial era. Continued CO_2 accumulation in the atmosphere puts Earth on a path to reaching 2 °C of warming later this century.

SPACE

Mission launch

NASA launched its Mars InSight spacecraft from Vandenberg Air Force Base in California on 5 May. The probe will record seismic activity on the red planet, including tremors known as marsquakes. Researchers hope to use the information to gauge the size of Mars's core, which could teach scientists about its density and composition, and the boundaries between the planet's core, crust and mantle. InSight will land in Elysium Planitia, a flat area near the Martian equator, on 26 November 2018. The international effort includes a seismometer built by a French team, and a heat-flow probe built by a German group.

FUNDING

Climate grants

Twenty-seven climate scientists from around the world are heading to work in France and Germany, as part of the second round of French President Emmanuel Macron's Make Our Planet Great initiative. Macron launched the €60-million (US$70-million) grant scheme last June in response to US President Donald Trump deciding to pull out of the Paris climate accord. Germany joined the grant scheme last September and added €15 million in government funding, which was matched by participating research organizations. The joint programme offers 4-year grants of up to €1.5 million for senior scientists and up to €1 million for early-career researchers. The second-round winners were announced on 2 May; 14 will work in France and 13 in Germany. Thirteen of the scientists currently work in the United States.

European science

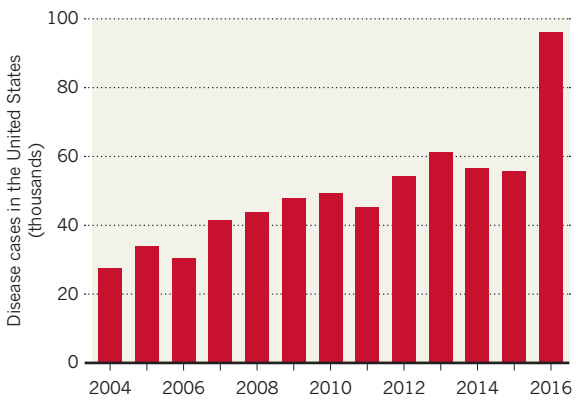
The European Commission has proposed a budget of €100 billion (US$120 billion) for its next major research-funding programme. Named Horizon Europe, it will last from 2021 to 2027, and follows the current programme, Horizon 2020. See page 150 for more.

TREND WATCH

Diseases spread by mosquitoes, ticks and fleas are rising in the United States, according to the US Centers for Disease Control and Prevention. Between 2004 and 2016, total yearly cases of a combination of 16 diseases tripled. Tickborne illnesses — Lyme disease, in particular — made up more than 75% of the nearly 643,000 cases during the whole period. Mosquito-borne infections, such as Zika and chikungunya, tend to spike with outbreaks, whereas the incidence of Lyme disease has risen gradually.

INSECT-BORNE DISEASES ON THE RISE

The number of infections from ticks, fleas and mosquitoes tripled in the United States between 2004 and 2016.



SOURCE: R. ROSENBERG *ET AL. MORB. MORTAL. WKLY REP.* **67**, 496–501 (2018).

NATURE.COM
For daily news updates see:
www.nature.com/news

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

NEWS IN FOCUS

CONSERVATION Yellowstone grizzlies could be hunted again **p.148**

BIOMEDICINE Viruses that fight cancer get fresh vote of confidence **p.150**

PHYSICS Universe's coldest lab will form window on quantum world **p.151**

CONSERVATION Seeking sanctuary on the high seas **p.154**



CERN

The pixel detector at CERN's CMS experiment records particles that emerge from collisions.

PARTICLE PHYSICS

Physicists turn to AI to cope with CERN collision deluge

Competition with cash rewards seeks better techniques for tracking particle debris.

BY DAVIDE CASTELVECCHI

Physicists at the world's leading atom smasher are calling for help. In the next decade, they plan to produce up to 20 times more particle collisions in the Large Hadron Collider (LHC) than they do now, but current detector systems aren't fit for the coming deluge. So last week, a group of LHC physicists teamed up with computer scientists to launch a competition to spur the development of artificial-intelligence (AI) techniques that can quickly sort through the debris of these collisions. Researchers hope that these methods will help the experiment's ultimate goal of revealing fundamental insights into the laws of nature.

At the LHC at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, two bunches of protons collide head-on inside each of the machine's detectors 40 million times a second. Every proton collision can produce thousands of new particles, which radiate from a collision point at the centre of each cathedral-sized detector. Millions of silicon sensors are arranged in onion-like layers and light up each time a particle crosses them, producing one pixel of information every time. Collisions are recorded only when they produce potentially interesting by-products. When they are, the detector takes a snapshot that might include hundreds of thousands of pixels from the piled-up debris of up to 20 different pairs of protons. (Because particles ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

move at or close to the speed of light, a detector cannot record a full movie of their motion.)

From this mess, the LHC's computers reconstruct tens of thousands of tracks in real time, before moving on to the next snapshot. "The name of the game is connecting the dots," says Jean-Roch Vlimant, a physicist at the California Institute of Technology in Pasadena who is a member of the collaboration that operates the CMS detector at the LHC.

After future planned upgrades, each snapshot will be expected to include particle debris from 200 proton collisions. Physicists currently use pattern-recognition algorithms to reconstruct the particles' tracks. Although these techniques would be able to work out the paths even after the upgrades, "the problem is, they are too slow", says Cécile Germain, a computer scientist at the University of Paris South in Orsay. Without major investment in new detector technologies, LHC physicists estimate, the collision rates will exceed the current capabilities by at least a factor of ten.

Researchers suspect that machine-learning algorithms could reconstruct the tracks more quickly. To help find the best solution, Vlimant and other LHC physicists teamed up with computer scientists, including Germain, to launch the TrackML challenge. For the next 3 months, data scientists will be able to download 400 gigabytes of simulated particle-collision data — the pixels produced by an idealized detector — and train their algorithms to reconstruct the tracks.

Participants will be evaluated on the accuracy with which they do this. The top three performers will receive cash prizes of US$12,000, $8,000 and $5,000.

PRIZE APPEAL

Such competitions have a long tradition in data science, and many young researchers take part to build up their CVs. "Getting well ranked in challenges is extremely important," says Germain. In a similar competition in 2014, teams competed to 'discover' the Higgs boson in a set of simulated data (the LHC discovered the Higgs, long predicted by theory, in 2012).

TrackML is "incomparably more difficult", says Germain. She thinks the winning technique might end up resembling those used by the program AlphaGo, which made history in 2016 when it beat a human champion at the complex game of Go. In particular, the methods might use reinforcement learning, in which an algorithm learns by trial and error on the basis of 'rewards' that it receives after each attempt.

Vlimant and other physicists are also beginning to consider more untested technologies, such as quantum computing. "It's not clear where we're going," says Vlimant, "but it looks like we have a good path." ■



Grizzly bears in the greater Yellowstone ecosystem could be targeted by hunters as early as September.

ROBBIE GEORGE/NGC/GETTY

CONSERVATION

Hunting proposal might threaten bears

Biologists argue that the plan could endanger grizzlies in iconic ecosystem that includes Yellowstone National Park.

BY GIORGIA GUGLIELMI

On 23 May, Wyoming officials will vote on whether to allow the hunting of up to 24 grizzly bears around Yellowstone National Park this September. The proposal has reignited controversy over whether or not this population has recovered from decades of hunting and habitat destruction — an issue central to the US government's decision to take the bears in the greater Yellowstone ecosystem off the endangered-species list in 2017.

Seventy-three scientists sent a letter to Wyoming Governor Matt Mead on 25 April, asking him to halt the hunt until independent experts can review data on the size of the grizzly (*Ursos arctos horribilis*) population in this area. They are concerned that government tallies overestimate the number of bears in this region, which spans roughly 80,000 square kilometres and is one of the largest continuous wilderness areas in the contiguous United States.

Critics challenge the federal government's methods of assessing whether the grizzly population is large enough to face a hunting season. The estimates might be too high because of several factors, says David Mattson, a wildlife researcher in Livingston, Montana, who retired from the US Geological Survey (USGS) in 2013. These include increased monitoring efforts in the past 30 years and better visibility of bears to aerial surveys because of shifts in where they look for food.

Wildlife scientist Frank van Manen, who leads the USGS Interagency Grizzly Bear Study Team (IGBST) in Bozeman, Montana, disagrees with critics of the government estimates. The IGBST collects population data using a range of methods, including aerial surveys and tagging bears, van Manen says, and the numbers from each method agree. The current population estimate of 718 bears is "extremely conservative", he says.

The Wyoming Game and Fish Department proposed the hunt in February on the basis of those population assessments, and gave the public until 30 April to comment on draft regulations. Under the proposal, hunters could kill up to 12 bears in the monitored region surrounding Yellowstone National Park, and a further 12 bears outside that area (but still in the greater Yellowstone ecosystem).

Mattson and the other researchers who wrote to the governor about the hunt listed several concerns in their letter. Some of the bear's food, such as cutthroat trout (*Oncorhynchus clarkii*), will probably become even scarcer in the future as a result of environmental changes, they say. This would push the animals to hunt livestock or look for food near

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

houses, increasing their run-ins with people, says Mattson. If the number of bears killed as a result of these conflicts increases, this would further shrink the population.

Even if the current population estimates are accurate, removing 24 animals through hunting could have detrimental effects, says Andrea Santarsiere, an attorney at the Center for Biological Diversity who is based in Victor, Idaho. In 2017, 56 bears died in the IGBST monitoring area as a result of natural causes or conflicts with people. "If the same amount dies this year, we could be looking at up to 80 bears removed from the population," Santarsiere says.

Killing females might pose even higher risks to the survival of these grizzlies, she says. The Wyoming proposal would allow the killing of up to two females in the IGBST monitoring area, but it doesn't cap how many females hunters can take outside this region. Females can carry up to four cubs at a time, Santarsiere says, "so killing one female could equal removing five bears from the population".

Van Manen says the hunting proposal won't pose a risk to the bear population. Only two hunters at a time would be allowed in the monitoring area, and the hunts would stop as soon as two females had been killed, he says.

Wyoming officials seem intent on allowing the hunt, says Louisa Willcox, a wildlife activist based in Livingston, Montana, who has been in contact with Wyoming's wildlife department. "It's extremely unlikely that the scientists' comments will make them pause." ■

FUNDING

Wellcome Trust vows to pull grants from harassers

UK charity launches policy to force institutions to report bullying or sexual misconduct.

BY HOLLY ELSE

One of the world's largest research-funding charities is cracking down on harassment and bullying. Scientists who have been sanctioned by their institutions could lose out on funding from the Wellcome Trust, under rules announced on 3 May.

It is the first major UK research funder to institute such a policy; the US National Science Foundation introduced similar rules earlier this year.

Wellcome's policy will come into force for new and existing grant applications on 1 June. It will apply to anyone already associated with a grant, including those whose projects are already under way. It gives Wellcome, a biomedical-research charity in London, the right to withhold funding from a researcher or to bar them from applying for future grants.

The policy also means that sanctions can be levied against institutions that fail to disclose details of such misconduct, do not investigate allegations in a timely and fair manner, or take inappropriate action. In extreme circumstances, sanctions could include suspending funding from an entire organization.

"Bullying and harassment are just plain wrong," says Alyson Fox, director of grants at the charity. These behaviours are harmful, and therefore affect the research that Wellcome funds, she adds. The policy "will give organizations notice that we are taking this extremely seriously".

The Wellcome Trust funded more than 900 grants, worth a total of more than £1 billion (US$1.4 billion), in 2017.

Under the new guidelines, Wellcome will require organizations that receive its grants to have clear policies that outline standards of acceptable behaviour by staff and procedures for responding to allegations of harassment and bullying.

The policy defines bullying as a misuse of power that can make people feel vulnerable, upset, humiliated, undermined or threatened. It says harassment is unwanted physical, verbal or non-verbal conduct that has the purpose or effect of violating someone else's dignity, or creating an intimidating, hostile, degrading, humiliating or offensive environment for them.

Six types of sanction can be applied to grant holders and Wellcome advisory committee members whose employers have investigated and upheld an allegation of bullying or harassment. They include removing researchers from grants, and banning them from supervising Wellcome-funded PhD students or submitting future grant applications.

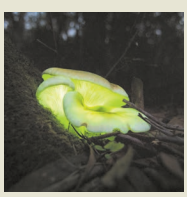
Institutions that do not abide by the policy risk being temporarily barred from applying for Wellcome grants. In extreme cases, they will have existing funded suspended.

Emma Chapman, an astrophysicist at Imperial College London and a member of the 1752 group, which lobbies against sexual misconduct in higher education, calls the harassment policy an "excellent step forward". However, she worries that it could lead universities to settle complaints informally to hide problems. The requirement to report only upheld allegations is understandable, Chapman adds, but it risks missing researchers who resign before an investigation is completed.

Philip Maini, a biological mathematician at the University of Oxford, UK, also questions how effective the policy will be. "If an institution has someone bringing in huge amounts of overhead and publishing in *Nature* and *Science*," Maini says, "are they really going to take action against them if they are a bully? I think not." ■

MORE ONLINE

IMAGES OF THE MONTH



PETAR B/REX/SHUTTERSTOCK

Billionaires, butterflies and brooding skies — April's top science images go.nature.com/2hzmtxy

TOP NEWS

- Water filter inspired by Alan Turing passes first test go.nature.com/2jgkjxr
- *Nature* announces new editor-in-chief go.nature.com/2hx8woa
- Report harassment or risk losing funding, says top UK science funder go.nature.com/2kha6ss

NATURE PODCAST

AI and our sense of place; liquid crystals deliver cargo; and depression in academia nature.com/nature/podcast

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

RESEARCH FUNDING

Europe's €100-billion science plan

Budget proposed for European Union's next big research-funding programme.

BY INGA VESPER

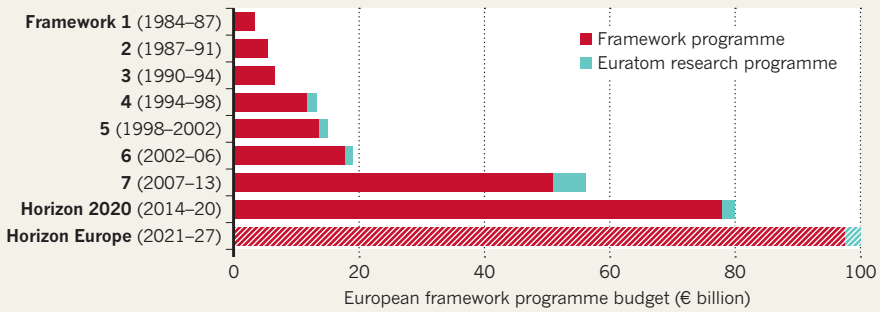
The European Union is planning to spend €100 billion (US$120 billion) on its next major research-funding programme, for 2021 to 2027 — a disappointment to some scientists and policy groups who had been hoping for up to 60% more. The budget does not include a contribution from the United Kingdom, whose departure from the bloc in 2019 is likely to shake up the distribution of funds among the remaining 27 EU countries.

The European Commission issued its opening budget proposal for Framework Programme Nine — newly named Horizon Europe, and the successor to the current programme, Horizon 2020 — on 2 May. The announcement marks the start of tough negotiations between the European Parliament and the Council of the European Union, which comprises government representatives from EU nations.

The proposed €100 billion is an increase on the €77-billion pot for Horizon 2020, which began in 2014 (see 'Europe's science spending'). However, a report by influential academic and industry experts, published last July, had urged a doubling of the budget for the next framework programme.

EUROPE'S SCIENCE SPENDING

The European Commission has proposed a €100-billion (US$120-billion) budget for Horizon Europe, the next instalment of its flagship research-funding programme, which will last from 2021 to 2027.



Euratom's budget is included as part of the framework's total funding, but its grants are distributed through a separate programme.

SOURCE: EUROPEAN COMMISSION

And in March, 13 science and higher-education organizations, including the European University Association (EUA), which represents more than 800 institutions, also demanded a €160-billion budget. "The increase is good, but it's not at the level we would consider suitable," says Enora Bennetot Pruvot, deputy director of governance, funding and policy at the EUA in Brussels.

"With the UK leaving the EU, we knew it was going to be difficult to get the €160 billion we would have liked to see," says Laura Keustermans, senior policy officer at the League of European Research Universities in Leuven, Belgium. The full proposal for Horizon Europe is expected in June. The framework is set to include funding for large, multidisciplinary 'missions' to tackle big societal questions; €10 billion has also been earmarked for research into food, agriculture and rural development. ■

BIOMEDICINE

Anti-cancer viruses take off

Encouraging study results and a handful of clinical trials spur interest in therapy approach.

BY HEIDI LEDFORD

Pharmaceutical giant Johnson & Johnson announced on 2 May that it would pay up to US$1 billion to acquire a company that makes cancer-killing viruses. The striking show of support for a still-unproven treatment is just the latest sign that industry and academics are warming to the approach.

In February, Merck, headquartered in Kenilworth, New Jersey, agreed to pay US$394 million to snatch up an Australian firm working on cancer-killing, or 'oncolytic', viruses. And in April, 300 people showed up for the oversubscribed International Oncolytic Virus Conference in Oxford, UK. When the conference launched in the early 2000s, there were only about 60 attendees. "They were very small meetings for these crazy people working with viruses," says Jean-Simon Diallo, a molecular biologist at the Ottawa Hospital Research Institute. "We've really seen a shift."

Diallo credits a couple of developments with igniting the field. One was a 2015 US Food and Drug Administration (FDA) decision to approve a modified herpes virus called talimogene laherparepvec (Imlygic) to treat some forms of melanoma. It was the first cancer-fighting virus to win regulatory support in the US market. Another development is emerging evidence — largely from animal studies — that the viruses might work better when administered in concert with therapies called checkpoint inhibitors, which boost immune responses against tumours.

People with cancer sometimes go into remission after contracting a viral infection.

"The intersection of these two events has really put some spice in the oncolytic-virus field," says Diallo. The checkpoint inhibitors in particular turned things around, he adds.

Researchers have been trying to develop cancer-fighting viruses for decades, hoping to capitalize on centuries-old observations that people with cancer sometimes go into remission after contracting a viral infection. That has spurred teams to develop a panoply of viruses that have passed through the gauntlet of a clinical trial.

Many of these trials have met with little success. Even Imlygic fell short of showing a statistically significant improvement in patient survival during a clinical trial (R. H. I. Andtbacka *et al. J. Clin. Oncol.* **33**, 2780–2788; 2015). Still, the results were enough to persuade the FDA to approve the therapy for melanomas that had resisted other treatments. That study

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

also sparked hope among researchers by showing that a virus injected directly into one tumour could rein in tumours elsewhere in the body.

DEATH BY ASSOCIATION

The viruses work by generating an immune response. After the virus infects and kills cancer cells, the immune system eliminates the virus and ends up also getting rid of the dead cancer cells. "The side effect from eliminating the virus is that the systemic immune system recognizes the cancer cells," says Tomoki Todo, a neurosurgeon at the University of Tokyo. "Then it starts to attack even those cancer cells that are not infected by the virus."

Scientists reasoned that bolstering such an immune response — for example, by using a checkpoint inhibitor — could amplify this indirect effect. These inhibitors sometimes send cancer into remission for years, but only for a fraction of people.

Studies in mice suggest that combining checkpoint inhibitors with cancer-killing viruses might boost that percentage. And in a small clinical trial involving 21 people with advanced melanoma, Imlygic, together with a checkpoint inhibitor, significantly shrank tumours in 62% of participants and wiped them out altogether in 33% (A. Ribas *et al. Cell* **170**, 1109–1119; 2017).

When researchers have combined checkpoint inhibitors with other treatments, they have met with mixed success. Some fervently anticipated combinations have failed in clinical trials. "Could the same thing happen with oncolytic viruses? Sure," says Dmitriy Zamarin, an oncologist at the Memorial Sloan Kettering Cancer Center in New York City.

But Zamarin and others are cautiously optimistic. Many checkpoint-inhibitor combinations target a specific protein, Zamarin notes, whereas oncolytic viruses provoke a much broader immune response that can target cancer in a number of different ways. "That gives us some comfort," he says. ■

QUANTUM PHYSICS

Universe's coolest lab set to open up quantum world

NASA's Cold Atom Laboratory will allow physicists to play with quantum phenomena like never before.

BY ELIZABETH GIBNEY

Quantum physicists are about to get their own playground in space. NASA's Cold Atom Laboratory, scheduled to launch to the International Space Station on 20 May, is set to be the coldest place in the known Universe. Researchers will use the lab to probe quantum phenomena that would be impossible to observe on Earth.

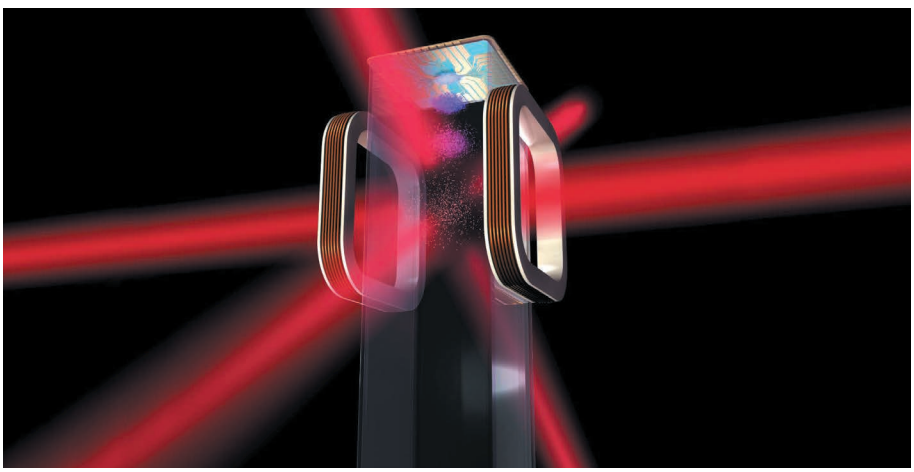
The US$83-million mission will be used to study quantum mechanics on the macroscopic scale by making a state of matter known as a Bose–Einstein condensate (BEC). These are clouds of hundreds of thousands of atoms that, when chilled to just above absolute zero, behave as waves that synchronize into a single quantum object. "Just being able to do these experiments in space I think is a huge accomplishment," says Kamal Oudrhiri, mission manager at the Jet Propulsion Laboratory (JPL) in Pasadena, California.

On Earth, gravitational forces usually disperse these condensates within a few seconds. The closest that BECs have come to being in space-like conditions is during brief stints in a research rocket, or falling from a drop tower over 9 seconds. But, floating on the space station, they should be able to exist for at least 10 seconds. That's long enough for them to be cooled to record-low temperatures — perhaps as little as 20 trillionths of a degree above absolute zero. That would be the coldest known temperature in the Universe, says Oudrhiri. Colder and long-lasting condensates will "push the frontiers of studying fundamental physics", says Gretchen Campbell, an atomic physicist at the US National Institute of Standards and Technology in Gaithersburg, Maryland. "It's something people have hoped for for almost 15 years."

DOWN-SIZED KIT

Real estate on the space station is at a premium, so engineers had to crunch down atomic-physics equipment that usually fills a large room into a cool-box-sized chest. The equipment will cool rubidium and potassium atoms by scattering laser light off the particles in all directions to slow them to almost a standstill. It will then use magnetic fields to trap the cloud. To create the condensate, other cooling techniques are used to push the cloud even closer to absolute zero — including creaming off the most energetic atoms using a radio-wave 'knife' and widening the trap to let the cloud expand.

Engineers also had to design shielding to protect the delicate condensates from interference from densely packed components and from Earth's varying magnetic field. Although astronauts will unpack and install the equipment, the experiments will run only while the team is asleep to minimize disruption from ▶



NASA/JPL-CALTECH

Instruments on NASA's Cold Atom Laboratory will cool atoms to near absolute zero (artist's concept).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

any movements.

The technology is simpler than initially intended, after a more-complex version of the lab developed a leak that affected the vacuum chamber and threatened to delay the project. So physicists will not yet be able to achieve their ultimate goal of performing space-based atom interferometry — a process that involves splitting the condensate's quantum wave in two and recombining the waves. The resulting interference patterns would allow scientists to analyse the effects of gravity with exquisite precision, as well as test whether condensates could be used as very sensitive rotation and gravity sensors. But the more-advanced kit should arrive by the end of 2019, says Robert Thompson, project scientist for the mission at JPL.

BUBBLES, RINGS AND WHIRLPOOLS

The current iteration will still allow new physics, says Thompson. Five teams are set to experiment with the lab; one plans to use radio waves and magnetic fields to trap the condensate in a bubble shape about 30 micrometres across — roughly half the width of a human hair. Quantum mechanics suggests that because the bubble is both thin and edgeless, the condensate should behave differently from when it is shaped as a disc or sphere on Earth. For instance, it might more readily form whirlpools known as vortices, says Courtney Lannert, a theoretical physicist at Smith College in Northampton, Massachusetts. On Earth, attempts to create bubbles always end up with bowl shapes as the fluid falls. "We can't get this shape at all unless we can get rid of gravity," she says.

A group led by Eric Cornell of the University of Colorado, Boulder, who won a Nobel prize in 2001 for co-discovering BECs, will try to create exotic loosely bound systems known as Efimov states. Named after Russian theoretical physicist Vitaly Efimov, who proposed their existence in 1970, these quantum states crop up where atoms would bind too weakly to join in pairs, but can form threesomes. These are similar to Borromean rings — rings linked in such a way that the system falls apart if any one ring is removed — and are of interest to nuclear physicists because they have parallels with rare and poorly understood three-particle nuclei made up of neutrons and protons. The team hopes to create the simplest Efimov state but also excited, bloated versions in which the atoms bind with each other despite being a bacterium-width apart. The group might also be able to make foursomes of such atoms, known as tetramers, says Maren Mossman, a physicist at Washington State University in Pullman.

"We can't get this shape at all unless we can get rid of gravity."

Atomic physicists will find the space-station set-up extraordinary for more-practical reasons, says Mossman. They are used to building their own equipment and tweaking experiments as they go. But with the Cold Atom Laboratory, many are sharing a facility for the first time, says Mossman, and they must run experiments through JPL researchers who are operating the facility from the ground. "Folks in particle physics have been doing this since the beginning. But it's so weird to us in atomic physics," she says.

The process worked out "better than most of us expected", says Thompson, who has been working to create such a facility since he joined JPL in 1997. He thinks that the current version is a step towards even more-complex atomic-physics labs in space. NASA is working with the German Aerospace Center (DLR) to build a facility called BECCAL (Bose–Einstein Condensate and Cold Atom Laboratory), he says. Many experiments on the space station already test the effects of low gravity, but for most, the extreme microgravity is "overkill", adds Thompson. "We are one of the experiments that will really highlight what the space station is capable of doing." ■

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

How to save the high seas

BY OLIVE HEFFERNAN



Vast expanses of the South Pacific, viewed here, are unprotected by law.

JAMES D. MORGAN/GETTY

AS THE UNITED NATIONS PREPARES A HISTORIC TREATY TO PROTECT THE OCEANS, SCIENTISTS HIGHLIGHT WHAT'S NEEDED FOR SUCCESS.

In the early fifteenth century, Portuguese sailors reached a becalmed part of the Atlantic Ocean, coated with mats of gold-brown seaweed. Under windless skies, their ships drifted idly with the currents. The sailors named the seaweed *Sargassum* — after its resemblance to a Portuguese plant — and the region eventually became known as the Sargasso Sea.

Initially thought to be an oceanic desert, this part of the Atlantic is now recognized as a watery rainforest. It is one of Earth's most rare and valuable marine ecosystems, so rich in nutrients that eels travel thousands of kilometres from rivers in Europe and the Americas to breed there.

But the Sargasso Sea is also one of the dirtiest and most damaged parts of the open ocean. The gyre of currents that bounds this shoreless sea entraps vast amounts of plastic waste, and fish stocks are declining in the now-busy shipping route.

Scientists want to conserve the Sargasso ecosystem, and ten governments have signed a non-binding pact to protect it. But their efforts are limited owing to a major gap in international policy. Like half of the planet, the Sargasso Sea doesn't fall under the control of any single nation. Countries can protect or exploit waters closer than 200 nautical miles (370 kilometres) to their shorelines, but everything outside these 'exclusive economic zones' is considered international waters: the high seas.

The high seas make up two-thirds of Earth's oceans, providing 90% of its available habitat for life and accounting for up to US$16 billion a year in fisheries catch. The oceans are also prime territory for the discovery of valuable mineral deposits, potent pharmaceuticals and oil and gas reserves. The United Nations Convention on the Law of the Sea (UNCLOS) regulates activities in international waters, including sea-bed mining and cable laying; a patchwork of 20 or so other organizations oversee aspects of international shipping and whaling, as well as fishing and conservation at the regional level. But no overarching treaty exists to protect biodiversity or conserve vulnerable ecosystems in the oceans.

Yet momentum is now building to protect the high seas. This September in New York City, negotiations begin on a United Nations treaty — which is likely to be an add-on to UNCLOS — to agree on how to safeguard this vast shared resource by setting aside areas for conservation and laying out rules for activities such as deep-sea mining. The treaty could also find ways to help all countries benefit from research into deep-sea species — including whether marine organisms' genes and proteins might form the basis of new drugs or materials — either financially or through technology transfer. The talks are being heralded as a Paris climate accord for the oceans: a vital opportunity to conserve the planet's least-explored realm. "We have a once in a lifetime chance to secure a treaty that will allow nations to manage activities on the high seas," says Lance Morgan, president of the non-profit Marine Conservation Institute in Seattle, Washington, which is focused on ocean protection.

The UN, regional fisheries organizations and non-profit agencies have already shortlisted numerous international marine regions that — like the Sargasso — deserve protection. But researchers are unsure whether politicians will heed scientific advice in choosing what to protect, and in making judgements about environmental impacts. Ahead of the negotiations, *Nature* lays out this guide to protecting the high seas, and the scientific debates at play.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Ocean havens

There's no shortage of ideas for marine protected areas (MPAs) on the high seas. UN organizations have listed dozens of vulnerable ecosystems, as have regional fisheries bodies and non-governmental organizations. This map highlights ten sites that showcase the diversity of ecosystems on the high seas and the range of threats they face. Data came from the Marine Conservation Institute, which has an interactive version at go.nature.com/2hlkked.

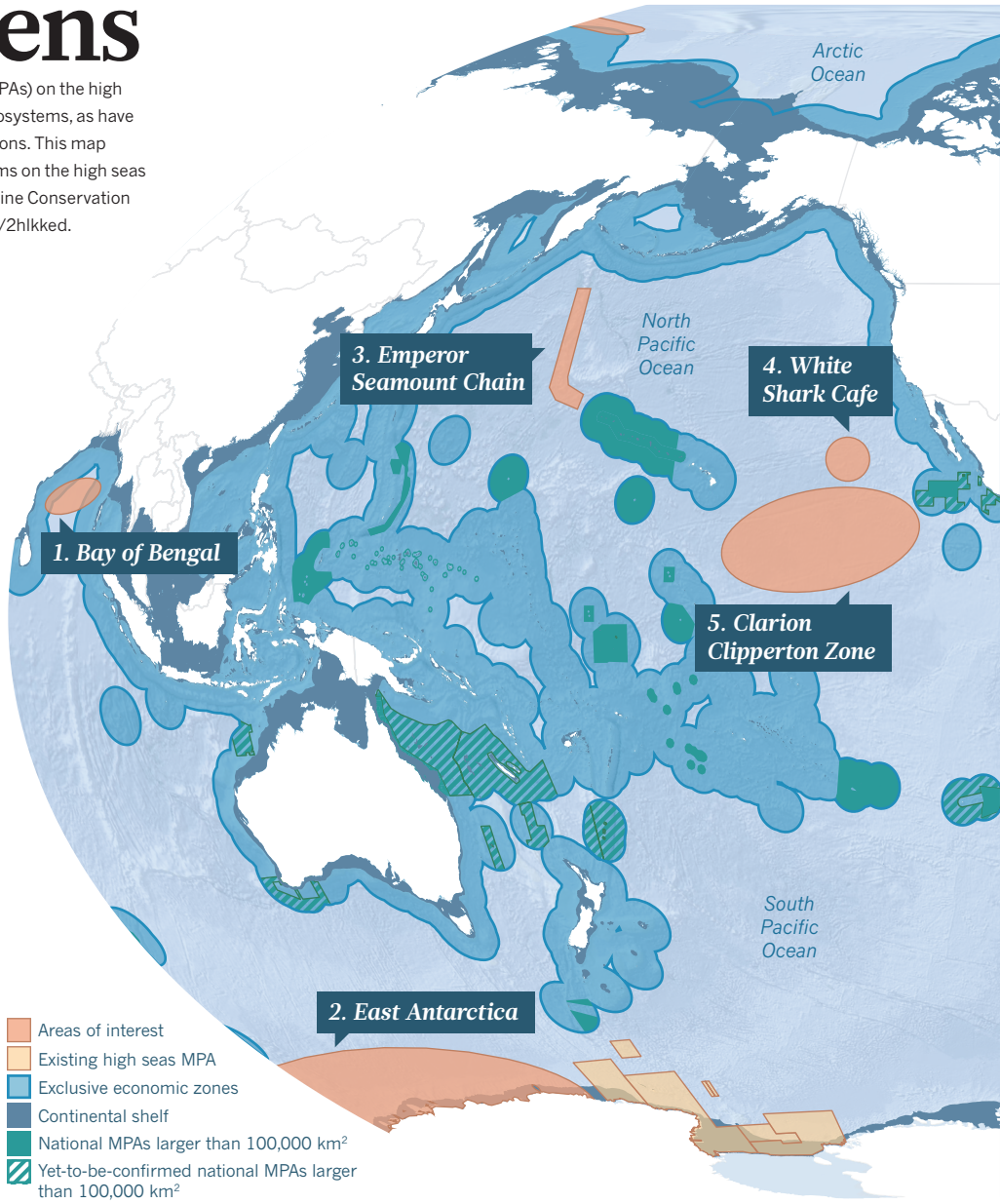
1. Dead zones. Pollutants from agricultural runoff can cause plankton blooms in the Bay of Bengal, a shallow, warm part of the Indian Ocean. The blooms suck up oxygen, leaving dead zones that total at least 60,000 square kilometres. Further runoff or a change in monsoons could cause huge-scale oxygen depletion, radically changing an ecosystem that provides jobs and food security to more than 100 million people.

2. First new MPA? East Antarctica, a relatively pristine part of the Southern Ocean that is home to penguins and corals, is a clear choice for a high-seas MPA. But China and Russia have interests in fishing krill here; in 2017, it was rejected as an MPA for the sixth consecutive year by the Commission for the Conservation of Antarctic Marine Living Resources.

3. Coral crunch. Between the Hawaiian and Aleutian islands, a chain of deep-sea volcanoes provides nutrient-rich waters for migrating albatrosses, whales and tuna. Corals and fish have been hit hard by trawling and are struggling to recover.

4. Shark cafe. Hundreds of great white sharks (*Carcharodon carcharias*) forage and breed here, in a region at risk from fishing and shipping. These sharks are a genetically distinct population and of higher concern even than other great whites; the species as a whole could number as few as 3,500 in the wild.

5. Sea-bed mining. Scattered on and below the sea bed are trillions of nodules — potato-sized, rock-like deposits rich in many valuable minerals. But the region also hosts rare marine species, including



HOW TO CORDON OFF THE OCEANS

A major focus of the treaty will be to agree a process to create marine protected areas (MPAs) — regions that are off-limits to at least some kinds of commercial activity. Established properly, MPAs can boost biodiversity in previously decimated regions. They can't stop plastics entering the ocean, or waters becoming hotter or more acidic, but they can increase the size and diversity of marine populations, giving them a greater resilience to these stressors.

Scientists say that at least 30% of the global ocean, distributed evenly between ocean ecosystems, should be cordoned off to avoid a mass extinction of marine life. On paper, almost 7% of the ocean is now protected: in the past 3 years, 13 of the world's largest MPAs, all more than 100,000 square kilometres in area, have been created in coastal waters — largely impelled by a UN goal to protect 10% of the ocean by 2020. In practice, however, these protections are often less than adequate. To be effective, MPAs need key traits: they must be 'no-take', or completely off-limits to commercial activity; have an area of at least 100 square kilometres; be permanent and physically isolated from their unprotected surroundings by deep water or sand; and have well-enforced protections. An analysis of 87 MPAs found that those with only one or two of these traits were ecologically indistinguishable from fished sites[1].

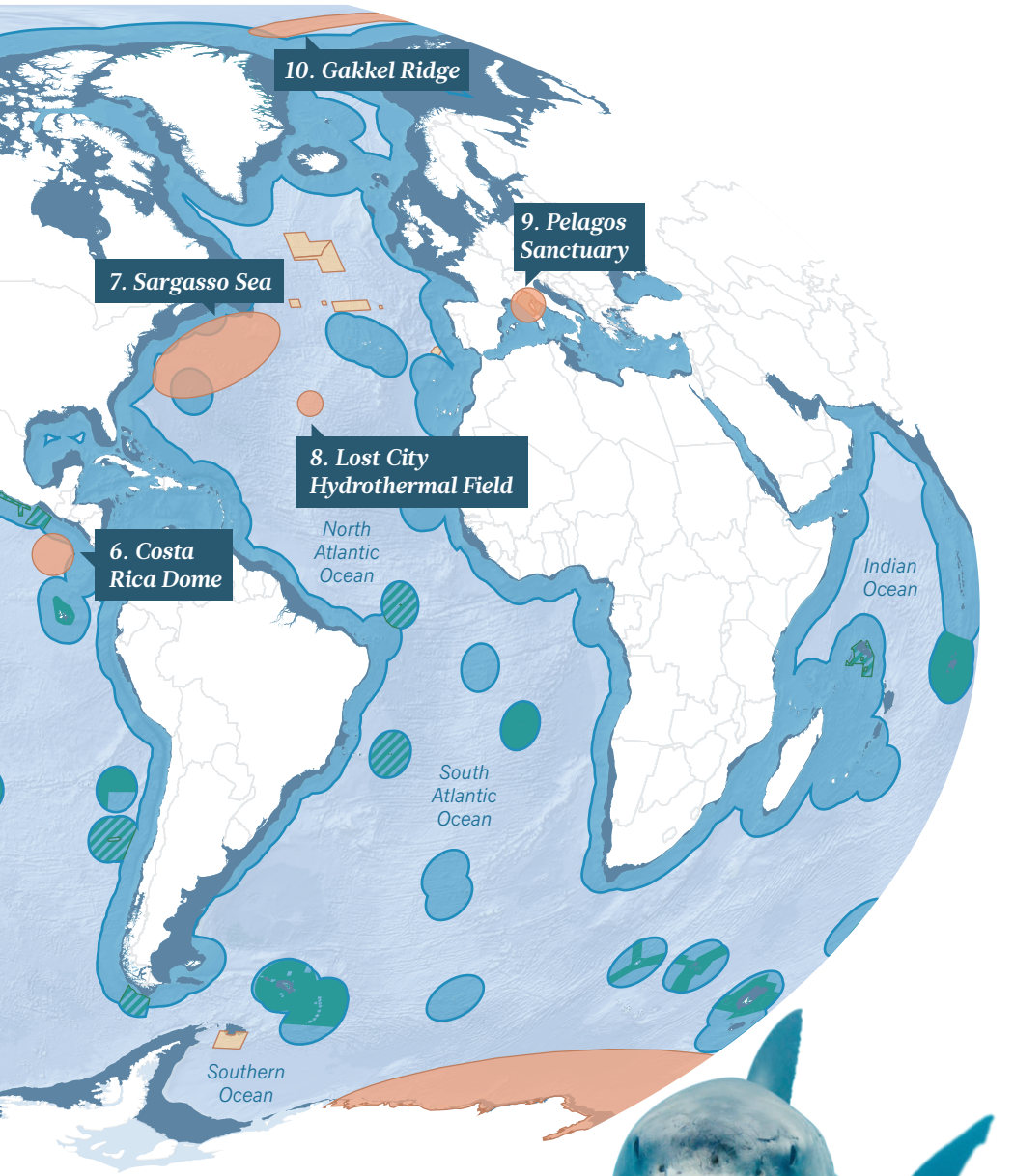
Many coastal MPAs allow for oil and gas exploration, shipping and fishing. Only 2% of the ocean is no-take, and these MPAs are mostly in deep tropical waters of little interest to industry, so do little to reduce overall exploitation of the ocean. As for the high seas, just 0.5% is off-limits to commercial exploitation. (Much of this is due to the largest international MPA, in the Ross Sea off Antarctica, which was created by a regional 25-nation council). "As is often the case closer to shore, there's a serious risk that high-seas MPAs will be sited in areas of low commercial interest," says Elizabeth De Santo, an environmental-management specialist at Franklin and Marshall College in Lancaster, Pennsylvania. How scientific advice on MPAs will feed into the UN treaty is yet to be decided.

MONITORING AND ENFORCEMENT

Once protected ocean areas have been agreed on, it's crucial to gather baseline data. A 2000–10 project called the Census of Marine Life provided much of what researchers know about life in the high seas, but the oceans have become warmer, more acidic and more heavily fished since then. This need for new data could stimulate a fresh period of discovery.

Monitoring breaches of protected areas is possible thanks to satellite technology. Global Fishing Watch (GFW), a satellite-based surveillance initiative that was launched in 2014 by the non-profit organizations SkyTruth and Oceana, together with Google, allows anyone with WiFi access to track fishers in real time, for instance. These data suggest that commercial fishing reaches more than half the ocean, spanning an area four times that covered by agriculture on land[2].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



DAVE FLEETHAM/GETTY



a species of ghost octopus that was discovered in 2016. The International Seabed Authority has issued 16 contracts to explore the area for minerals. Scientists say at least one-third of the zone should be off-limits to mining, with controls in place where it is permitted.

6. Dynamic dome. Strong winds drive currents that force cold, nutrient-rich waters to well up from the deep to just below the surface. Iconic ocean species come here, including mahi-mahi, billfish, sharks, squid, cetaceans and endangered sea turtles. But this 'thermic dome' shifts its position, and only seasonally occurs on the high seas, so it is challenging to protect.

7. Marine rainforest. The Sargasso region is one of 37 EBSAs, or 'Ecologically or Biologically Significant Marine Areas' on the high seas. The UN designation identifies the regions as important to healthy ocean function but does not protect them.

8. Hydrothermal field. Discovered in 2000, the 'Lost City' system could give clues to the necessary precursors for life on Earth. At a depth of 800 metres, this acidic, hot ecosystem extends for about 400 metres along the top of an underwater mountain known as the Atlantis Massif. The United Nations Educational, Scientific and Cultural Organization has proposed a 20-kilometre buffer zone.

9. Ineffective sanctuary. This refuge, regarded as the first ever high-seas MPA, was created in 1999 to protect the many cetacean species that visit its waters. But the sanctuary lacks management and has had little effect. If expanded and implemented properly, it could provide refuge for bluefin tuna, sharks and swordfish.

10. Oil and gas. This 1,800-kilometre mountain chain hosts active volcanoes, hydrothermal vents and unique creatures such as eyeless shrimp (*Rimicaris exoculata*), which could be vulnerable to shipping and oil and gas exploration as the Arctic warms.

But prosecuting regulatory breaches is a political issue — and at the will of individual nations. An analysis of hundreds of coastal MPAs has found that staffing and budgets are the strongest predictors of whether an MPA will have a conservation impact[3]. The ecological effects of MPAs with enough staff to patrol and monitor activity within the reserve were nearly three times greater than were those of MPAs with inadequate capacity, researchers found.

ENVIRONMENTAL ASSESSMENTS

On land and in coastal waters, new commercial activities have to undergo an 'environmental impact assessment' or EIA, to weigh up any benefits against potential harm to local wildlife.

On the high seas, only some activities are regulated in this way. It wasn't until 2006 that bottom trawling — a highly destructive fishing method — needed an EIA. Before that, it destroyed deep-sea corals. Even now, mid-water fisheries, open-water farming and rocket launches (which dump waste at sea) do not need to consider potential environmental harm. Scientists want to see new commercial activities on the high seas tightly regulated. Deep-sea mining, in particular, is likely to cause a flashpoint at the UN talks.

JUST 0.5% OF THE HIGH SEAS IS OFF-LIMITS TO COMMERCIAL EXPLOITATION.

The International Seabed Authority, established by UNCLOS, has approved 29 exploration licences for companies such as Lockheed-Martin to do surveys, mostly along ridges in the oceans and at hydrothermal vents. It is now developing regulations for what EIAs mining companies would have to conduct.

Cindy Van Dover, a deep-sea biologist at Duke University in Durham, North Carolina, says that scientists are concerned these regulations will not be strict enough. One unresolved issue is whether companies can mine active hydrothermal vent sites, which support large, diverse biological communities. "We're arguing that we should protect active hydrothermal vents," she says. Scientists don't yet know whether these communities can recover from mining. "What we've learned from bottom trawling is that the recovery time, particularly for complex habitats like deep-water corals, can potentially be hundreds of years," says Paul Snelgrove, a deep-sea biologist at the Memorial University in St Johns, Canada. "I think we have to accept that science will not be the only deciding factor, but we certainly hope it will be one of the major considerations," he says. ■

Olive Heffernan *is a science writer in Dublin.*

1. Edgar, G. J. *et al. Nature* **506**, 216–220 (2014).
2. Kroodsma, D. A. *et al. Science* **359**, 904–908 (2018).
3. Gill, D. A. *et al. Nature* **543**, 665–669 (2017).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

COMMENT

MENTAL HEALTH "Through the darkest days, I never hated the science" **p.160**



ENERGY A history of the pioneers of power from steam to nuclear **p.162**

PHYSICS Richard Feynman's books, from hilarious to revolutionary **p.164**

DEVELOPMENT New national funds for research in Kenya and Ghana **p.166**

CLAUDIA DAUT/REUTERS



A man waits for treatment at a hospital in Havana, where a drug to treat diabetic foot ulcers was shown to be effective in a clinical trial.

When will clinical trials finally reflect diversity?

An analysis of drug studies shows that most participants are white, even though trials are being done in more countries, reveal **Todd C. Knepper** and **Howard L. McLeod**.

Since the late 1990s, the number of countries contributing to the clinical-trial data used by the US Food and Drug Administration (FDA) to approve drugs has almost doubled (see 'Going global'). Yet this global expansion of study locations has not been accompanied by an equivalent increase in the racial diversity of people enrolled. In 1997, 92% of the participants in these trials were white; in 2014, we found that this figure was still nearly 86%.

A growing body of literature indicates that the effectiveness of a drug, the likelihood of it causing side effects and the nature of those effects can all vary between people of different ancestry[1]. And funders and researchers have repeatedly said that clinical trials should include more participants from ethnic minorities. Indeed, 25 years ago the US National Institutes of Health Revitalization Act called for more people from ethnic minorities to be included in clinical trials.

In our view, drug developers should capitalize on the global expansion of clinical-trial locations to design studies that represent more of the world's population.

PROBING THE DATA

To understand which populations provide the drug safety and efficacy information ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

used by the FDA, we reviewed the approvals made by the agency at five time points from 1997 to 2014. For each approval, we documented the reported race of the people involved in the clinical trial and the country where the trial was conducted. (All were pivotal efficacy trials, generally the most definitive demonstrations of a medicine's efficacy and safety.) We focused only on the categories 'white', 'black', 'Asian' and 'other', because the reporting for these was uniform across countries. The data for ethnicities, such as Hispanic or non-Hispanic people, were confounded by reporting inconsistencies and were therefore omitted from the analysis.

We focused on treatments for heart disease, cancer and disorders of the central nervous system (CNS); 41% of the drug approvals made over the period 1997 to 2014 were for these common global health problems[2]. We assessed FDA approvals because many countries follow the agency's lead when it comes to their own regulatory decisions.

During the five years we assessed (1997, 2004, 2009, 2012 and 2014), 81 drugs for heart disease, CNS disease and cancer won FDA approval (see Supplementary information) on the basis of clinical trials involving nearly 150,000 people. Twenty-nine countries contributed clinical-trial data in every year evaluated. For eastern Europe, Asia-Pacific, Latin America and the Caribbean, representation increased over the study period (see 'Going global'). We classified regions according to the United Nations Regional Groups.

Over the five time points, the racial make-up of the clinical trials stayed relatively stable (see 'Going global'). The median percentage of African and African American participants per trial ranged from 1.8% to 3.5%. For Asian participants, the range was 0% to 7%; for any group unspecified or not described as white, black or Asian, it was 1.4% to 3.4%. For context, according to the US census, 72.7% of the US population was non-Hispanic white in 1997 and by 2014, this figure was 62.2%. Also, as of 2015, around 75% of the global population lived in Asia or Africa[3]. Of course, our analysis has limitations. The ideal mix of race and ethnicity will vary from country to country; we compared the trial demographics to the US population because we used trial data evaluated by the FDA.

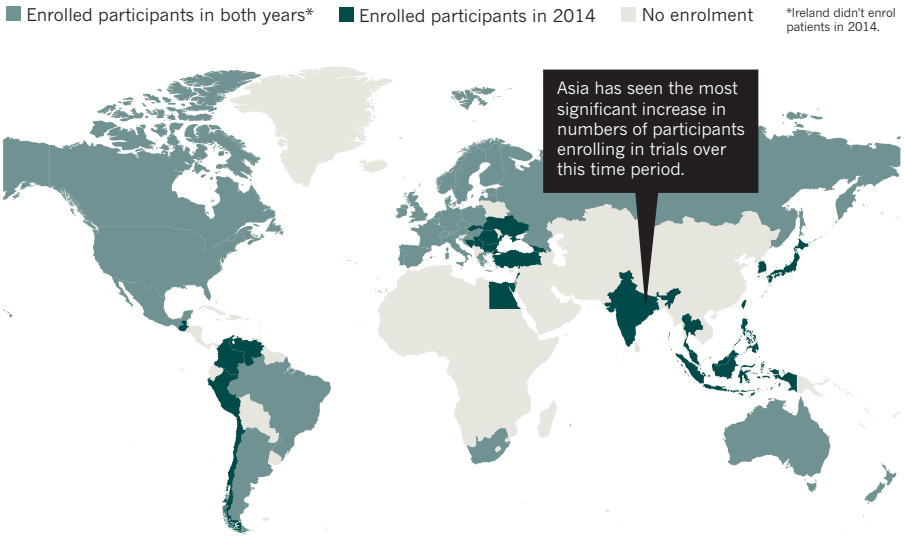
SHIFTING LOCATIONS

There are various possible reasons why the diversity of trial participants is not increasing with the expansion of countries contributing to trial data. It is likely that fewer people are enrolling in the trials or fewer trials are being conducted in those countries that have begun contributing data more recently.

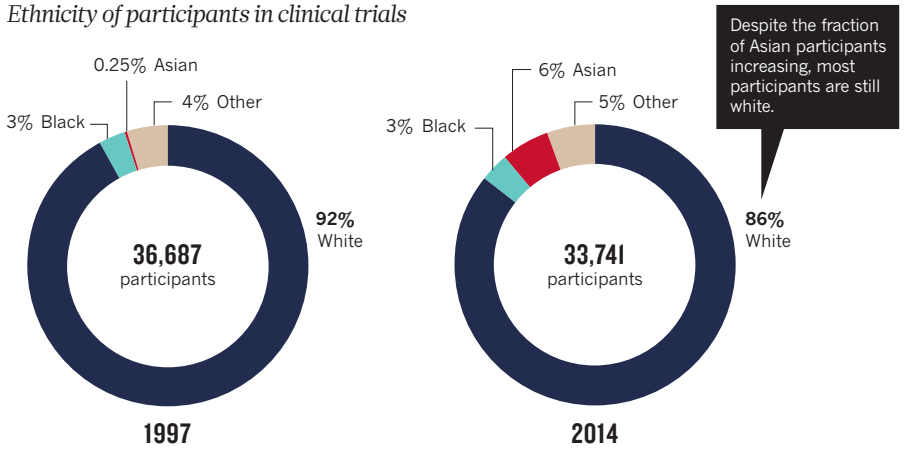
GOING GLOBAL

The number of countries contributing to the clinical-trial data that the US Food and Drug Administration uses to approve drugs almost doubled between 1997 and 2014, from 32 to 57. The racial make-up of the trials was little changed.

Countries contributing to clinical-trial data



Ethnicity of participants in clinical trials



SOURCE: US FOOD AND DRUG ADMINISTRATION

None of the conversations we have had with drug developers suggests that the goal of including greater population diversity in clinical trials is driving the change in where trials are being conducted.

We think that the sharp increase in the number of countries contributing to clinical-trial data (from 32 in 1997 to 57 in 2014) has been driven by two main factors.

The first is a growing shortage of people from developed economies who are eligible for clinical trials. Wealthy countries are better able to adopt new therapies into standard care than are poor countries. So many patients in the United States and western Europe now have access to a variety of medications outside a clinical-trial setting. The scarcity of people with a particular condition who are not already taking medication — 'treatment naive' patients — has prompted drug developers to start recruiting overseas.

The second factor is lower costs. In North America and western Europe, hospitals and other care centres tend to charge drug developers much more for hosting clinical trials than do equivalent institutions in eastern Europe and Asia. Staff expenses also tend to be lower in less-wealthy countries and studies can be completed more quickly.

DRUG EFFICACY

Many studies indicate that the likelihood, nature and severity of side effects from a medication can differ between populations[1].

For example, the antiplatelet drug clopidogrel reduces a person's likelihood of having a heart attack or a stroke after some heart procedures. Genome-wide-association and other studies have revealed that people with certain genetic variants of the *CYP2C19* gene, which encodes an enzyme that activates the drug, might need a different and more expensive therapy[4]. And people of

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Asian heritage are three times more likely to produce poorly functioning CYP2C19 enzymes than are people of white heritage[5,6].

Similarly, around 20 years ago, genomic and other studies, predominantly conducted in white populations, identified genetic markers that indicated whether people with cancer would have dangerously low counts of white blood cells after being treated with the autoimmune medicines 6-mercaptopurine (6-MP) or azathioprine[7]. The markers were variants in a gene called *TPMT*. Studies subsequently conducted in Asian and South American patients receiving these drugs identified variants in *NUDT15* as the key predictor of whether the drug is toxic to a person's white blood cells[8,9]. An important patient-safety marker had been missed for nearly two decades owing to a lack of comprehensive genomic testing.

MOVING FORWARD

Various efforts are under way to raise awareness about the importance of population diversity in clinical trials, and to improve the quality of the data that are collected on race and ethnicity (see 'Documenting diversity'). Too much of what we know about population differences in drug responses is anecdotal or stems from observations of a handful of patients. Most clinical trials do not enrol enough patients from diverse populations to provide definitive guidance.

It is obviously not practical to try to conduct drug trials in every recognized patient population in the world. But we think that developers are missing an opportunity to capitalize on the increase in the number of countries hosting clinical trials. Researchers should be designing studies — and choosing study sites — to make clinical trials more informative about the safety and efficacy of drugs for as much of the global population as possible.

This could mean conducting clinical trials in particular places. Or it could mean increasing the numbers of patients recruited to certain trials. Another option could be to conduct 'ethnobridging' studies. Here a drug that has been assessed in a major clinical trial is then tested on a smaller number of people from a population of interest to gain insights about side effects, appropriate dosing levels and so on for that population. Japan uses this approach a lot: medications deemed safe and effective from global trials are subsequently tested on Japanese patients[10].

Regulatory authorities could do more. Mechanisms for homing in on 'special populations' in drug development are already in place. For instance, the FDA frequently requires developers to conduct a further trial focused on a certain age group if a large proportion of those likely to use the drug in question will be, for example, older than 75. The agency and other regulators could apply similar mechanisms to race or ethnicity. Regulators could also provide guidance for manufacturers on how genetic-ancestry information could be used to expand the assurances for clinicians and patients around drug safety.

Ensuring that more populations are represented in clinical trials might also require that developers, regulators and others address some of the social barriers to enrolment.

Within any one country, some populations might, for cultural or historical reasons, prefer not to engage in a scientific study. Or they might have limited access to the medical centres where clinical trials are being conducted. Currently, some investigators are working to address this, for instance by trying to build trust with communities on their turf. As far as we know, no such efforts are being pursued at scale by regulators or by pharmaceutical companies.

"Most clinical trials do not enrol enough patients from diverse populations to provide definitive guidance."

The ideal would be to match treatments to individual patients on the basis of their genetic, proteomic or other profiles. In theory, that would remove the need to even consider race or ethnicity. With this goal in mind, identifying the individual differences that influence the risks and benefits of a medication should be a component of clinical trials.

For now, however, investigators must at least exploit the fact that clinical trials are being done in many more countries than they were two decades ago — and strive to obtain a more complete picture of disease and how to treat it, to the benefit of all. ■

DOCUMENTING DIVERSITY

Regulatory efforts

Various international endeavours have attempted to provide guidance on how to consider and document population diversity in clinical trials.

The International Conference on Harmonisation (ICH) has produced reports on this issue. The ICH brings together the regulatory authorities of Europe, Japan and the United States with experts from the pharmaceutical industry to create standards for the conduct of clinical trials. The ICH E5 report sets out how to evaluate the impact of ethnicity on a medicine's efficacy and safety. The ICH E17 (released in 2017) offers guidance on planning and designing clinical trials conducted across multiple locations[11].

Similarly, in 2012, it became a requirement in US law that the Food and Drug Administration report on the extent to which demographic subgroups are represented in clinical studies for new drugs, medical devices and biologics such as monoclonal antibodies. The agency has produced reports summarizing these data, including the analysis *Global Participation in Clinical Trials* using data from 2015 to 2016 — an evaluation that is largely consistent with our findings.

Building on these reports, the FDA updated its guidance on how to collect race and ethnicity data in clinical trials. And in 2014, it began releasing Drug Trials Snapshots, public information on participants in the clinical trials that have supported the agency's approval of new drugs (including people's sex, race, age and so on).

These are steps in the right direction. But they have not yet driven a sufficient increase in the diversity of populations represented in clinical trials. T.C.K. & H.L.M.

Todd C. Knepper *is a personalized medicine specialist and* **Howard L. McLeod** *is chair of the Department of Individualized Cancer Medicine at DeBartolo Family Personalized Medicine Institute, Moffitt Comprehensive Cancer Center, Tampa, Florida 33612, USA.*
e-mails: todd.knepper@moffitt.org; howard.mcleod@moffitt.org

1. Ramamoorthy, A., Pacanowski, M. A., Bull, J. & Zhang, L. *Clin. Pharmacol. Ther.* **97**, 263–273 (2015).
2. Mullard, A. *Nature Rev. Drug Discov.* **14**, 77–81 (2015).
3. United Nations *Population and Vital Statistics Report* (United Nations, 2014); go.nature.com/2HUz7JA
4. Wang, Y. *et al. J. Am. Med. Assoc.* **316**, 70–78 (2016).
5. Shu, Y. & Zhou, H. H. *Acta Pharmacol. Sin.* **21**, 193–199 (2000).
6. Scott, S. A. *et al. Clin Pharmacol. Ther.* **94**, 317–323 (2013).
7. McLeod, H. L., Krynetski, E. Y., Relling, M. V. & Evans, W. E. *Leukemia* **14**, 567–572 (2000).
8. Moriyama, T. *et al. Nature Genet.* **48**, 367–373 (2016).
9. Knepper, T. C. & McLeod, H. L. *J. Clin. Oncol.* **33**, 1230–1231 (2015).
10. Shirotani, M., Suwa, T., Kurokawa, T. & Chiba, K. *J. Clin. Pharmacol.* **54**, 438–445 (2014).
11. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *General Principles For Planning And Design Of Multi-Regional Clinical Trials E17* (ICH, 2017); go.nature.com/2jik8bz

Supplementary information accompanies this article online (see go.nature.com/2I224Wm).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



ILLUSTRATION BY NEIL WEBB @ DEBUT ART

You are not alone

Twenty years on, **Dave Reay** speaks out about the depression that almost sunk his PhD, and the lifelines that saved him.

A black dog. That's how Winston Churchill described depression. A dark companion that lies quiet in the good times and is your master in the bad. Its arrival is hard to predict, just as its departure seems impossible until it has happened. Depression is as varied as the people who experience it, as shifting in form as a murmuration. Yet its flavour is unmistakable.

It is twenty years since my own black dog stalked away. Twenty years since debilitating depression saturated my every thought and act. Survivors are forever wary that the dog will return. That it will overcome the new defences, leap the redoubts. It's always out there somewhere, skulking; reminding us to watch our step.

Mental illness affects around a quarter of us at some point in our lives[1], with more than 300 million people battling depression worldwide[2] and academics at greater risk than many[3]. The pressure to publish, make an impact, win funding, chase tenure, engage the public, shine on social media and influence policy, combined with failing experiments, lone working and rigid hierarchies, are all threats to our well-being. Yet our job is a thing of beauty, too. To explore new ideas, to grapple with the unknown, and ultimately to understand: that is a privilege.

As scientists, we talk, a lot. To our students and peers, to policymakers and funders. We become experts in telling the world about our work and opinions, but we risk not truly hearing ourselves or those around us.

My own two decades of silence were borne of fear. Fear of ridicule and mistrust, of overt hostility and covert isolation. Talking to friends, sharing memories of the dark times, has tempered these fears. Although speaking openly about my illness still makes me anxious, it is nothing to the waking nightmare lived by my student self.

If I could speak to him now, I'd tell him not to be afraid, to trust in people. I'd tell him how wonderful the warmth of those he spoke to will feel, how sharing the pain of his

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

mental illness would diminish its effects, not him. I would hug him and whisper: "You are not alone."

I tell this story for anyone who is silenced by the stigma of depression. I tell it to thank the wonderful people who gave me a voice, and then listened. I tell this story because I can, because I survived.

Early one year in the mid-1990s my undergraduate road was nearing its end. Three years of marine biology, substance abuse, mild poverty and falling in love. By October, I was a doctoral student 500 kilometres away from family and the love of my life.

It was natural to be scared. New people, new place, unknown rules and expectations. On the tall stool at my carefully demarcated square of lab bench I sat each day, trying to fathom the instructions of my supervisor. The research was on impacts of warming in the Southern Ocean — growing algae in the lab for starters, maybe a trip south later on. It was a masquerade from the start. The techniques and equipment were alien, my nodding response to instruction, an act. I tried to mimic the easy confidence of the other students, to fit in. My supervisor and his group must have assumed I had some idea of what to do. I hadn't. Not a clue.

Batch after batch of my growing medium went bad. Each morning, returning to the lab from my bedsit, the large glass bottles would have the unmistakable milkiness of contamination. Hours of weighing, mixing and sterilization would follow as I flushed with embarrassment and tried to suppress a mounting fear. Fear of being reprimanded. Fear that I was just too stupid. Fear that everyone would find out.

The days merged into weeks; the failures continued. My supervisor called me to his office. He was unimpressed. I needed to do better, work harder, get it right. Walking back to my bench that day, a black dog walked with me. When exactly it had arrived I can't say. I was 21. My ramshackle mental defences had been crumbling for a while. I'll never know what tore the breach. It may have been an admonishment from a lab technician or just one more cloudy culture bottle. Whatever had splintered the final defence, depression was my new master.

It is amazing how we can maintain a facade of normality while behind it a maelstrom of disintegrating sanity roars. By Christmas, I was thinner and quieter, but still me. Time with family and fiancée kept the black dog subdued. It waited.

Instead of seeking help, I counted down the remaining days of holiday like a condemned man. Sometimes you can see deep depression coming for you, slamming doors of escape, tightening the orbits of desperation. Deciding I would kill myself was a relief. I wrongly felt that it was the only door left open. One place the black dog could not follow.

Saying goodbye to my family when I returned to university that January was brutal. My mask could not withstand the warmth of those people I so loved. Through the sobbing, fear drove me on, piloting my body through the long journey back to my term-time lodgings.

Alone in the house I ran a hot bath, drank whisky without tasting it and selected the sharpest kitchen knife. I sat in that bath for an age. The knife hurt on my flesh, whisky or no. I wept and retched and wept some more. Too afraid to cut deep, anxious about what people would think, about the pain. I drained the bath and went to bed.

"Even through the darkest days, gripping my lab bench to suppress the convulsions within, I never hated the science."

The months that followed were a fluctuating grey-scale of near-normality and despair. Depression is exhausting. Even eating and washing felt irrelevant to my own contracting existence. Some days were doable, others I spent lying balled and wide-eyed, waiting for the night. When it arrived, sleep delivered only confused marathons of lost time.

My second attempt at suicide came just before Easter and was longer in the planning, the means gleaned from the Dick Francis mystery *Comeback* and obtained from the lab's chemical store. An injection, some nausea, a short blackout, that's all.

Throughout this period I self-medicated with alcohol, and cannabis when it was available; anything for oblivion. I didn't seek help, therapy, or prescription drugs, nor did I quit my PhD. I survived long enough for help to find me, eventually.

By the summer, life settled. Research, relationships, even my supervisor: all the quakes that had brought down the walls of my mind were still. I laid plans for work on an icebreaker in the South Atlantic that coming winter. It was exciting, an opportunity I had dreamed of, and yet still one in which I pictured my own destruction: deep inside, thoughts of suicide lingered, a secret obsession. A vision of slipping unseen into a remote sea, all a terrible accident, a tragic loss.

It was people who saved me. They still do. A warm word from a friend, a joke and a moan with a colleague. Day by day, piece by piece, the people around me unknowingly brought my mind back to health. One man in particular — a beautiful human and can-do lab technician named Paul Beckwith — helped me to trust in others, to share and to belong.

In December, just over a year after the start of my PhD, I joined the research ship near the Falkland Islands and my convalescence accelerated. I sipped daily on a tonic of big-hearted, inspirational and funny scientists. Being stuck together in floating isolation for months could have been torment. Instead it was my salvation.

The cloudy jars and experimental despair were replaced by thriving baths of algae, and geeky thrills as the microbes responded to artificially induced future climates. Even through the darkest days, gripping my lab bench to suppress the convulsions within, I never hated the science. Now that I could finally understand how to do it, I loved it.

The savage beauty of the South Atlantic, its sparkling wildlife and scientists, will always be a touchstone for me. A month after joining the ship and at the end of a long night shift, I stood alone on its rolling deck. Dawn broke, and tears streamed down my face. Tears of joy and release. The black dog was gone.

The two decades since have sewn their own patchwork of light and dark. Deaths of those close to me, including the beautiful Paul, have called my black dog close again at times. But, for me, it is kept at bay by lucky brain chemistry, and the warmth and understanding of my wife (that love of my life), and of our children and friends. Not everyone is so fortunate.

As the supervisor of dozens of students and staff, empathy is useful, listening is vital. Just a chat, making time for people; it can mean the world. We are mentors, not therapists, but our university communities are under enormous stress. Insecure jobs and mounting debt, endless metrics and poor management — all are risks to our mental health in the edifice of corporatization that our higher-education system has become.

Awareness is growing, with supervisor training, peer-support networks and counselling services now commonplace. More importantly, the stigma of mental health issues is fading. Whether it's to a friend or loved one, a therapist or colleague, those afflicted by mental illness must feel secure in speaking out. Silence was for too long the enemy of my own healing. Next time that black dog comes close, it will be met with a roar. ■

Dave Reay *is professor of carbon management and education, School of Geosciences, University of Edinburgh, UK.*
e-mail: david.reay@ed.ac.uk

1. Kessler, R. C. *et al. Epidemiol. Psichiatr. Soc.* **18**, 23–33 (2009).
2. World Health Organization *Depression and Other Common Mental Disorders: Global Health Estimates* (WHO, 2017); available at http://go.nature.com/2ksfnhg
3. Royal Society & Wellcome Trust *Understanding Mental Health in the Research Environment: A Rapid Evidence Assessment* (Royal Society, Wellcome Trust, 2017); available at http://go.nature.com/2p8fq8r

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



HERITAGE IMAGES/GETTY

George and Robert Stephenson's Rocket won an 1829 competition to find the engine for the world's first passenger railway.

ENERGY

Titans of power

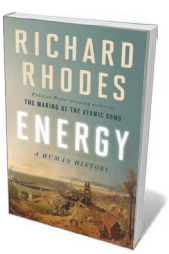
Roger Fouquet critiques Richard Rhodes's study of energy pioneers.

Richard Rhodes is renowned for his magisterial four-volume series on nuclear weaponry, starting with *The Making of the Atomic Bomb* (1987). Now, he sets his sights on an even broader landscape. *Energy* sweeps from the end of the sixteenth century to the golden era of energy before the 1970s oil shocks, charting the challenges of engineering power, light, warmth and low-polluting technologies. Focusing on the West's rise to energy supremacy through mastery of technologies from steam to nuclear power, Rhodes explores the choices that led us to a present threatened by climate change.

The bulk of his history centres on creative individuals, rather than economics or broader social and cultural implications of their work. As he shows, most of these engineers and inventors were not hermit geniuses. They often collaborated with peers, absorbed ideas and benefited (in 'shoulders of giants' mode) from earlier inventions. In re-enacting the build-up to pivotal events, Rhodes brings us into the eureka moments behind them.

He captures both the limitations of technologies and the improvements that inventors offered. In the 1760s, for instance, master instrument-maker James Watt boosted the efficiency of steam engines by separating the working cylinder from the condenser. Rhodes also deftly portrays intellectual debates. We read the late-eighteenth-century correspondence of electrical experimentalists Luigi Galvani and Alessandro Volta, which revolved around giving shocks to frogs' legs. And he relates the nineteenth-century battle of alternating and direct currents between Thomas Edison and George Westinghouse.

Rhodes portrays innovators and entrepreneurs as obsessively dedicated, often at the expense of health and prosperity. Engineer Richard Trevithick, inventor of the high-pressure steam engine, abandoned his family in 1816 to go to Peru, where his machines were deployed to dredge silver mines. He made and lost fortunes, and returned to see others (such as George Stephenson and his son Robert) develop the first steam railways. He died a pauper. Society, Rhodes shows, rarely repaid its debt to the era's technologists.



Energy: A Human History
RICHARD RHODES
Simon & Schuster (2018)

Political intrigue figures large. Inventors squabbled over patents, or lobbied government through intermediaries to introduce laws protecting their interests. Thomas Savery's experience is fascinating. The Enlightenment engineer's connections enabled him to extend the patent on his early steam pump by 21 years, through the 1699 Fire Engine Act. As a result, in 1712, Thomas Newcomen — who had designed a more powerful engine using atmospheric pressure — was forced to go into partnership with Savery. The engines could pump water out of deep mines, solving a major problem.

Rhodes also stresses the harm of energy exploitation, such as the devastating toll wreaked on cetacean populations by the whaling industry, in pursuit of oil for lighting. He highlights disasters, from London's Great Smog of 1952, which killed thousands, to the ongoing threat of climate change.

Rhodes's discussions about environmental damage are welcome, and his tour

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

of bold innovators is well guided. But he fails to mention the women who helped to revolutionize the field — from discoverer of nuclear fission Lise Meitner to solar-power pioneer Mária Telkes. Nor does he devote much space to renewable energy. His history is uneven in other ways: the latter third of the book is a critique of the anti-nuclear lobby that, in my view, skews the overall message.

Rhodes believes that protesters have stymied our transition to nuclear power, as has the adoption of a "linear no-threshold" model of the biological effects of radiation, which proposes that even low levels are potentially harmful. Enough evidence is in place to question this model. But it is interesting how Rhodes repeatedly reminds us that the energy industry has a history of denying damage in the interest of profit, yet he seems to believe the nuclear industry immune to this practice.

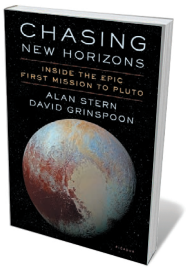
Rhodes argues for the commercial viability of nuclear technology. He calls the US project to build an early pressurized-water reactor in the 1950s "a godsend"; by 1954, he claims, "it was already competitive with non-nuclear power in Western Europe and Japan". In my opinion, this ignores the full costs. No nuclear power station has ever won a competitive tender. Private companies still refuse to invest in the technology without massive government support. With a lack of significant investment in Europe and the United States, today's nuclear power industry is focused in China, where subsidy is a given and cost is less of an issue. I feel that, instead of blaming protesters, Rhodes should acknowledge that the failure of nuclear power is down to the fact that the economics haven't worked.

From this perspective, the book's prime conclusion is highly optimistic. Rhodes asserts that humanity will be able to produce power on demand for hundreds of years more. He is concerned about climate change, but believes that low-carbon energy sources (including renewables and nuclear power) will solve the problem. This may well be true; perhaps the international experiment ITER and its successors will manage to provide cost-effective nuclear fusion. But innovators will need not only to create new kit. They will also need to spur new policy levers and behavioural patterns — and even social and political systems.

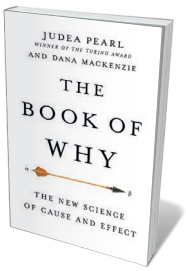
Energy is packed with good stories but, ultimately, I see it as like a Thomas Savery engine: it generates a lot of heat, but loses steam. A scholar of Rhodes's stature should have offered a deeper understanding of our struggle to improve our energy capacity — one of our most pressing global challenges. ■

Roger Fouquet *is associate professorial research fellow at the Grantham Research Institute on Climate Change and the Environment at the London School of Economics and Political Science.*
e-mail: r.fouquet@lse.ac.uk

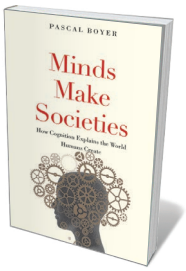
Books in brief



Chasing New Horizons
Alan Stern and David Grinspoon PICADOR *(2018)*
It is the farthest-flung planetary mission yet. In July 2015, NASA spacecraft New Horizons zipped past the Pluto system, 5 billion kilometres from Earth, at a cool 52,000 kilometres an hour. The stunning portraits it snapped range from Pluto's heart-shaped Tombaugh Regio and towering spires of methane ice to its zoo of moons. Here, mission leader Alan Stern and astrobiologist David Grinspoon trace decades of painstaking preparation, and look to the craft's future in the Kuiper belt and beyond — possibly heading into interstellar space. An exhilarating trek into the "wild black yonder".



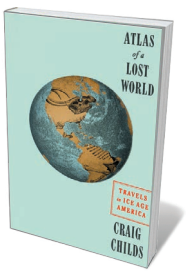
The Book of Why
Judea Pearl and Dana Mackenzie BASIC *(2018)*
"Correlation is not causation." That scientific refrain has had social consequences — not least, the long debate over links between tobacco and lung cancer. Judea Pearl proposes a radical mathematical solution. A prominent researcher in artificial intelligence (AI), Pearl teams up with writer Dana Mackenzie to explore the science of causation, now bearing fruit in biology, medicine, social science and AI through tools such as the "causal inference engine". The inevitably chewy narrative is leavened by the contributions of half-forgotten researchers such as psychologist Barbara Burks.



Minds Make Societies
Pascal Boyer YALE UNIVERSITY PRESS *(2018)*
Why do people believe absurdities, tolerate demagogues and cooperate (or not)? To decipher the complexities of human society, evolutionary psychologist Pascal Boyer draws from genetics, economics and biology. His incisive investigation of "six problems in search of a new science", from group conflict to information, thus integrates an impressive range of research and attendant insights. He reveals, for instance, how ethnic strife is fed by a belief that national welfare is a zero-sum game, and why splashy wedding ceremonies cement pair-bonding. An elegant voice in a clamorous field.

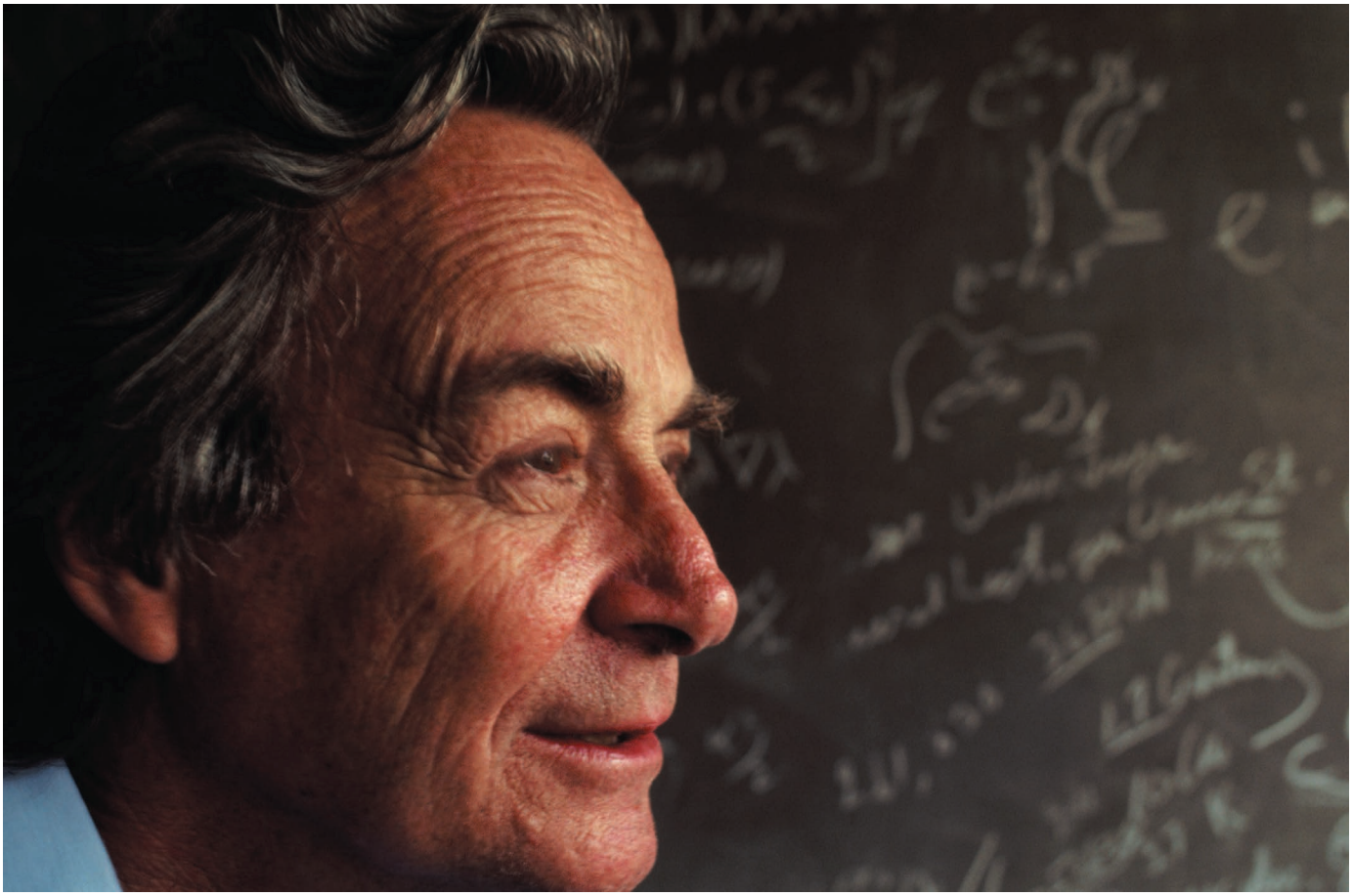


Milk!
Mark Kurlansky BLOOMSBURY *(2018)*
Best-selling author Mark Kurlansky follows up *Cod* (1997) and *Salt* (2002) with another zestful exploration of one foodstuff — milk — through history and a range of lenses. A human staple for 10,000 years, milk was the first food to be investigated in a modern lab. It is now the most globally regulated. Kurlansky keeps up a cracking pace on a tour that covers classical geographer Strabo griping about the Celts' milk consumption; the disease-generating dairies of nineteenth-century New York City; lactose intolerance in China; and 126 recipes for everything from ghee to syllabub.



Atlas of a Lost World
Craig Childs PANTHEON *(2018)*
What was life like for the early humans who poured over the Bering land bridge to North America more than 20,000 years ago? On evidence, beyond brutal terrain and climate shifts, it involved harrowing games of hide-and-seek with colossal fauna such as the short-faced bear *Arctodus simus*. In this evocative scientific travelogue, Craig Childs enters that time and space, braiding together his treks along prehistoric pathways from Alaska to the south with palaeontological findings from coprolites to mastodon bones, and meetings with remarkable scientists. **Barbara Kiser**

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



KEVIN FLEMING/CORBIS VIA GETTY

Richard Feynman shared the 1965 Nobel Prize in Physics.

THEORETICAL PHYSICS

Feynman at 100

Paul Halpern celebrates the oeuvre of a brilliant, original scientist on his centenary.

A pre-eminent twentieth-century physicist and a Nobel laureate: Richard Feynman was certainly those. He was also much more. As the centenary of his birth rolls around on 11 May, a look at his scientific and cultural legacy recalls his restless multidimensionality. His popular books shattered readers' preconceptions of scientists as lab-coated nerds and replaced them with a hipper image of a wild non-conformist; his scholarly tomes introduced researchers to revolutionary methods of grappling with modern physics.

Feynman was a master conjuror of physics. A mathematical whizz with exceptional intuition, he seemed to pull solutions out of thin air. He crafted a lexicon for particle interactions: iconic squiggles, loops and lines now known as Feynman diagrams (see D. Cressey *Nature* **489**, 207; 2012). His Nobel-prizewinning work on quantum electrodynamics included methods that even he saw as a sleight-of-hand for removing infinite terms from calculations. Yet, his results — equivalent to more systematic, rigorously expounded mathematical techniques independently proposed by co-laureates Julian Schwinger and Sin-Itiro Tomonaga — matched atomic-physics data beautifully.

Decades before his 1965 Nobel Prize, Feynman was already a legend of the Manhattan Project at Los Alamos, New Mexico, where he helped develop the atomic bomb. His colleagues and supervisors, including scientific director J. Robert Oppenheimer and head of the theoretical division Hans Bethe, were stunned by his computational abilities.

Then there was the playfulness. His pranks, including safe-cracking and sneaking through security fences, and his passion for playing the bongos, were arguably as memorable as his science. Feynman loved telling stories about himself and observing the reaction — the more stunned, amused or horrified the better. In the 1980s, his friend and fellow drummer Ralph Leighton collected some of them in two bestselling volumes: *Surely You're Joking, Mr. Feynman!* (1985) and *What Do You Care What Other People Think?* (1988).

PLAYFUL SPIRIT

In the first, Feynman flaunted his rough edges and eccentricities, and much of the book is hilarious. (One story sees him bungling a sprinkler experiment in the cyclotron laboratory at Princeton University in New Jersey, shattering glass tubes and flooding the space with water.) The chronicles of Feynman's Los Alamos days are exceptionally funny, such as his removal of the secret contents of physicist Edward Teller's locked desk drawer after Teller told him it was impenetrable. However, the book shows its age in disturbingly sexist sections such as "You just ask them?", about his predatory behaviour towards women.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

His relationships with women were complicated. In *What Do You Care*, he explained how he encouraged his younger sister Joan, now an acclaimed astrophysicist, to go into science. And he recounted how many of his attitudes had been shaped by his love for his first wife, Arline, who died of tuberculosis in 1945, a few years after they married. Those stories were mostly written during the final stages of Feynman's life, after he had undergone treatment for cancer. By that point he had undoubtedly become more cognizant of his legacy.

Indeed, beneath his clown's guise, Feynman was a sensitive man, suffering from both early grief and considerable anguish about the atomic weapons he had helped engender. These demons stymied his research from the end of the Second World War until 1947, when findings reported at the Shelter Island Conference in New York helped to spark his Nobel-prizewinning work. At the meeting, physicist Willis Lamb presented evidence of a discrepancy between predictions for certain energy levels of the hydrogen atom, calculated using the Dirac equation, and experimental results obtained using microwaves to excite the atom. Inspired by Bethe's quick calculation of this Lamb shift, Feynman developed techniques to solve that problem and beyond, to the widest range of quantum electrodynamic interactions between charged particles.

GOLDEN YEARS

The next two years proved extremely productive in disseminating his extraordinary methods for calculating interactions in particle physics. Feynman published seminal papers such as 'Classical electrodynamics in terms of direct interparticle action' (J. A. Wheeler and R. P. Feynman *Rev. Mod. Phys.* **21**, 425–433; 1949) and 'Space-time approach to quantum electrodynamics' (R. P. Feynman *Phys. Rev.* **76**, 769–789; 1949). These build on each other like revelations in a Sherlock Holmes story — with the last supplying a complete resolution of how electrons interact using photons. Through his diagrams, which surprisingly depict positively charged positrons as electrons moving backward in time, Feynman portrayed the gamut of possible modes of interaction, each contributing to the total picture in a weighted tally called "sum over histories". In his brilliant vision, quantum reality is a cloud-like smear of particles' possible paths, as if a commuter from Essex to London could travel the entire route by train, coach, car and bicycle simultaneously.

Feynman published two excellent primers introducing these quantum methods. *Quantum Mechanics and Path Integrals* (1965; co-authored with Albert Hibbs) begins with one of his favourite apparatuses, the double-slit experiment, and quickly moves into path integration and other advanced methods. The second, *QED* (1985; based on a lecture series) is more accessible — explaining the same theory using diagrams and examples.



CALTECH ARCHIVES

Richard Feynman lecturing at California State University, Long Beach, in 1979.

Starting in the 1950s, Feynman ventured into many other areas of theoretical physics: superfluids, superconductivity, gravitation and the constituents of protons and neutrons, which he called partons. Two later papers that made a huge impact were his model of the weak interaction, 'Theory of the Fermi interaction' (R. P. Feynman and M. Gell-Mann *Phys. Rev.* **109**, 193; 1958), and his proposal for quantum computation, 'Simulating physics with computers' (R. P. Feynman *Int. J. Theor. Phys.* **21**, 467–488; 1982). In each, he tinkered with variations until he reached definitive conclusions.

"Feynman's books urge us to explore the world with open-minded inquisitiveness."

Feynman was also a renowned educator. Those lucky enough to have attended his lectures have the best sense of how his agile mind operated. He taught a first-year course at the California Institute of Technology (Caltech) in Pasadena: 'Physics X', in which students would ask him anything and he'd think on his feet. Feynman loved to astound, and often refused to provide solutions, to spur students on intellectually. The careful notes of attendees have been published as books and articles, bolstering his reputation as a master lecturer. One such from Caltech was the three-volume *The Feynman Lectures on Physics* (1964). Another, 1959's *The Theory of Fundamental Processes*, is based on notes taken by Peter Carruthers and Michael Nauenberg, two students at Cornell University in Ithaca, New York, when Feynman was a visiting lecturer there in 1958. Nauenberg told me how, during the first lecture, Feynman walked in, glanced at the blackboard, wildly erased the equations on it and declared that they would all learn the whole of physics from scratch. Within a few weeks, the course proceeded from elementary quantum mechanics to Feynman's rules for particle-physics calculations. *The Character of Physical Law* (1967), another of Feynman's works, emerged from lectures he delivered at Cornell six years later.

Feynman's books urge us to explore the world with open-minded inquisitiveness, as if encountering it for the first time. He worked from the idea that all of us could aspire to take the same mental leaps as him. But, of course, not every ambitious young magician can be a Harry Houdini. Whereas other educators might try to coddle those who couldn't keep up, Feynman never relented. The essence of his philosophy was to find something that you can do well, and put your heart and soul into it. If not physics, then another passion — bongos, perhaps. ■

Paul Halpern *is a professor of physics at the University of the Sciences in Philadelphia, Pennsylvania. He is the author of 15 science books, most recently* The Quantum Labyrinth *and* Einstein's Dice and Schrödinger's Cat.
e-mail: p.halper@usciences.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Correspondence

Sequencing risk to antibody patents

Patenting claims for genetic sequences of potential therapeutic interest could be called into question if they match those in natural DNA. Next-generation sequencing (NGS) combined with single-B-cell isolation technologies could reveal millions of new human antibody sequences (see S. Rajan *et al. Commun. Biol.* **1**, 5; 2018; B. Wang *et al. Nature Biotechnol.* **36**, 152–155; 2018).

NGS data are currently deposited at the Sequence Read Archive (see go.nature.com/2hhafpd). Although proper sequence annotations and analysis pipelines are still crucially needed (for searching the NGS data archive, for example), these data stand to reveal a wealth of information about human antibody repertoires and immunogenetics.

However, patent applications for huge numbers of such newly discovered antibodies could be doomed — given the 2013 US Supreme Court ruling that naturally occurring DNA segments are non-patentable.

Prabakaran Ponraj *Intrexon Corporation, Germantown, Maryland, USA.*
pponraj@intrexon.com
P.P. declares competing financial interests: see go.nature.com/2kwzwxt

Bitcoin mining built on shifting sands

I disagree with Stuart Wimbush's arguments on the wealth generated during cryptocurrency mining (*Nature* **555**, 443; 2018). Attributing value to the work itself rather than to its results works in a theoretical economics framework, but risks failing to capture the complexity of what wealth actually means.

Many would argue that wealth is defined by social stability and by access to food, education and health care, rather than just by market values (see, for example, R. Costanza *Nature* **556**, 300–301; 2018). Living in an intact environment is also a form of wealth. The presupposition that cryptocurrency mining generates wealth, purely by converting vast amounts of energy into the results of otherwise purposeless algorithms, is in my view misguided — akin to claiming that multiplying the values of all state currencies by ten would make the world population ten times wealthier.

Wimbush's assertion that US$275 billion in wealth was generated through Bitcoin and Ethereum currency mining throughout 2017 overlooks their continuing and substantial devaluation. Bitcoin dropped by 49% against the US dollar in the first quarter of this year alone. 'Wealth' is merely a fluctuating number, which can be wiped out in a matter of weeks. I perceive the environmental damage that remains (see S. Foteinis *Nature* **554**, 169; 2018) as a negative global-wealth coefficient.

Maciej Kawecki *Swiss Federal Laboratories for Materials Science and Technology, Dübendorf, Switzerland.*
maciej.kawecki@empa.ch

Secure energy, food and water in north

Permafrost thawing in the warming Arctic and sub-Arctic is likely to cause widespread land subsidence, damaging homes and infrastructure and disrupting energy, food and water supplies. It could be particularly hard for developed countries around the North Pole to ensure that remote communities have access to these essentials by 2030, in accordance with the United Nations sustainable development goals.

One way to help would be for governments to encourage adoption of renewable energy in the circumpolar north. This could power water-treatment facilities and greenhouses, thereby enhancing water and food security, creating jobs and making communities self-sufficient.

Many in the far north of Alaska, Greenland and the Canadian and Russian Arctic are not connected to centralized grid systems (L. Mortensen *et al. Polar Geogr.* **40**, 163–185; 2017). Distribution of renewable energy would therefore depend on off-grid solutions such as community electricity generators, or on setting up and reconfiguring local microgrids.

As more financial and technological support is provided to build community-scale renewable-energy facilities, subsidies on fossil fuels in the circumpolar north could gradually be phased out.

Alex C. Y. Yeung *University of British Columbia, Vancouver, Canada.*
alex.yeung@alumni.ubc.ca

Kenya and Ghana's new science funding

Most African countries still do not prioritize research and development, despite its potential for resolving problems such as the continent's food security and its heavy disease burden. The governments of Kenya and Ghana aim to change this by setting up national funding programmes to achieve high-quality research output.

Kenya's government has committed 2% of the country's gross domestic product (GDP) to research. Its national research fund offers competitive grants to researchers in academia and the private sector. These financed 389 graduate-student fellowships in 2016 (see go.nature.com/2ji6lso) and 158 multidisciplinary or multi-institute research proposals (see go.nature.com/2rdduxv). In addition, 20 infrastructure development grants were awarded — to create centres for research into stem cells, cancer, fishing and food safety, for example.

Ghana's government also intends to start its own national research fund, with a view to raising the money it spends on research to 1% of GDP. Last year it proposed investing the equivalent of US$50 million, pending parliamentary approval (see go.nature.com/2w140tn). It also plans to set up a National Science, Technology and Innovation Fund that is accessible to researchers and all potential inventors and innovators (see go.nature.com/2w2jjgl).

Proper management will allow these initiatives to serve as a model for progress in other African countries. Government officials, grant managers and scientists must ensure, for example, that funding is awarded on merit and released on time.

Thomas K. Karikari *The Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden.*
Patrick Amoateng *University of Ghana, Accra, Ghana.*
ohenekakari@gmail.com

Fourier's Egyptian transformation

You mark the 250th anniversary of the birth of Joseph Fourier by exploring the past and present ramifications of his celebrated mathematical transform (*Nature* **555**, 413; 2018). Also noteworthy is his expertise in Egyptology.

As you mention, Fourier joined Napoleon Bonaparte's Egyptian expedition as an administrator in 1798. He quickly became an accomplished scholar of Egyptology. It is said that in 1802 he showed a copy of the Rosetta Stone inscription to the young Jean-François Champollion (1790–1832), declaring it impossible to decipher the hieroglyphs it bore. Twenty years later, Champollion announced his success in decoding some of these hieroglyphs (see also A. Robinson *Nature* **450**, 793–794; 2007).

Andreas Otte *Offenburg University, Germany.*
andreas.otte@hs-offenburg.de

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

BIOTECHNOLOGY

Scissors become sensors

Two bacterial Cas proteins cleave nucleic acids indiscriminately after binding to specific target sequences. This property has now been harnessed to create highly sensitive, portable diagnostic tools for detecting viruses at low cost.

D. DEWRAN KOCAK & CHARLES A. GERSBACH

People infected with a virus often harbour very low numbers of circulating viral particles, making the infection difficult to detect. For example, during one outbreak involving Zika virus, infected individuals harboured as little as one virus copy per microlitre of blood serum[1]. These small numbers pose a large problem for medical diagnostics. However, researchers might in the future be armed with molecular tools whose sensitivity has been refined by evolution. Writing in *Science*, three groups[2–4] report a new class of diagnostic tool, created by repurposing nucleic-acid sensing systems found in bacteria.

CRISPR–Cas systems are adaptive immune systems that protect bacteria from viral attack by functioning as nucleic-acid sensors. Nucleic-acid sequences from viruses that have previously invaded a bacterium are stored in the cell's genome, and are used by Cas enzymes as a guide with which to recognize and bind to the same sequence if the virus invades again. Incoming viral nucleic acids are then cleaved through sequence-specific interactions.

One particular CRISPR system, derived from the bacterium *Streptococcus pyogenes*, has been adapted for use as a genome-editing tool that is revolutionizing biomedical research. However, there are many other types of CRISPR system, each of which has specific characteristics. The three current studies use CRISPR systems that have an unusual property: after engaging a target nucleic acid, they cleave other nearby nucleic acids indiscriminately, a process called collateral activity. The studies use this activity as a signal amplifier, whereby the cutting of many arbitrary nucleic acids serves as a beacon for the detection of one specific molecule.

In the first of the current papers, Chen *et al.*[2] examined a type V CRISPR system that uses Cas12a proteins to target viral DNA. The authors discovered that Cas12a has collateral activity, indiscriminately cleaving single-stranded (ss) DNA after binding to its specific double-stranded (ds) DNA target. The researchers took advantage of this property to create a diagnostic tool for detecting viruses.

In the authors' diagnostic, Cas12a guide sequences are designed to match particular targets in the viral DNA. After binding a viral target, Cas12a cleaves a 'reporter' construct — an ssDNA that is tagged at one end with a fluorophore molecule and at the other with a quencher molecule. Before cleavage, the fluorescence emitted by the fluorophore is captured by the quencher. However, after the ssDNA has been cleaved by Cas12a, fluorescence can be readily detected (Fig. 1a). Chen and colleagues boosted this fluorescent signal using a process called recombinase polymerase amplification (RPA) to amplify the target viral DNA, generating many more copies of the target sequence than are present in the original sample.

The group's system, which they name DETECTR (DNA endonuclease targeted CRISPR *trans* reporter), can detect target DNA at extremely low concentrations — as little as one molecule per microlitre of sample analysed. The authors also showed that DETECTR could discern different strains of human papillomavirus in crude DNA extracts from patient samples.

In the second paper, Gootenberg *et al.*[3] expanded on previous work[5–7] to create a diagnostic called SHERLOCKv2 (specific high sensitivity enzymatic reporter unlocking version 2). The original version of SHERLOCK works similarly to DETECTR, but uses Cas13, which binds to and indiscriminately cleaves RNA, rather than DNA[5]. SHERLOCK is not limited to RNA detection, however — a transcription step can be included in the RPA process to produce RNA from viral DNA, enabling DNA detection by proxy[6].

Gootenberg and colleagues improved SHERLOCK in several ways. First, they screened 17 Cas13 family members and thoroughly characterized their activity. The authors discovered that a subset of Cas13 variants

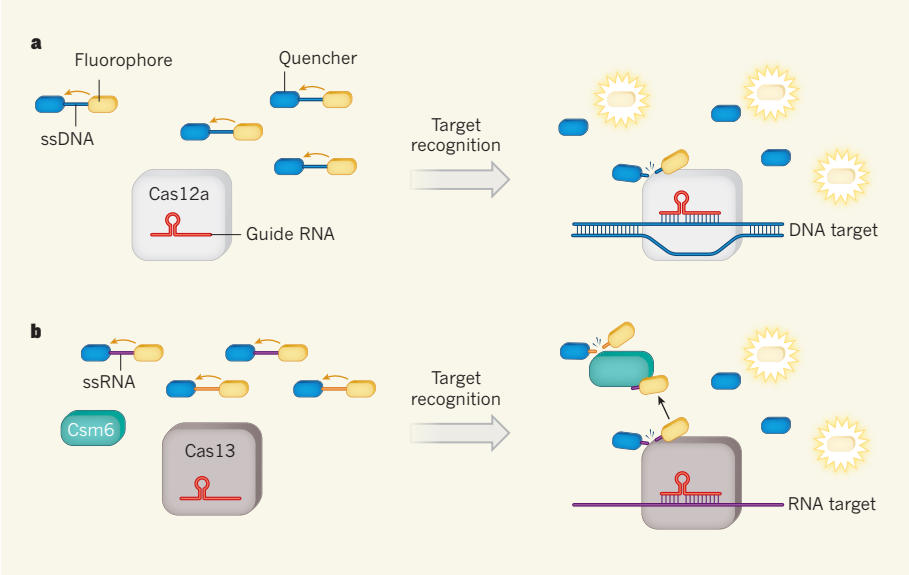


Figure 1 | Repurposing CRISPR–Cas to produce virus sensors. **a**, Chen *et al.*[2] have developed a virus sensor called DETECTR, which has three parts: a Cas12a enzyme; an RNA sequence designed to guide Cas12a to a matching DNA sequence in a virus of interest; and a single-stranded (ss) DNA construct tagged with 'fluorophore' and 'quencher' molecules. In the absence of the target virus sequence, Cas12a is inactive and emission produced by the fluorophore is captured by the quencher (yellow arrow), preventing fluorescence. When Cas12a recognizes its target, it cleaves the ssDNA construct, separating the quencher from the fluorophore to generate fluorescence. **b**, Gootenberg *et al.*[3] and Myhrvold *et al.*[4] developed tools based on a system called SHERLOCK (ref. 6) that is similar to DETECTR. SHERLOCK uses Cas13, which targets viral RNA and cleaves ssRNA constructs. Gootenberg *et al.* designed two ssRNA constructs, one cleaved by Cas13 and one by another Cas enzyme, Csm6. Csm6 is activated by Cas13 cleavage products, and serves to amplify the fluorescent signal and improve the reaction kinetics.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

target mutually exclusive sequences. By using four such Cas13 proteins and designing a specific reporter for each, the authors created a SHERLOCKv2 system that could detect up to four viruses at once. Furthermore, by scaling up the RPA step, the group detected viral DNA sequences at concentrations as low as two copies per millilitre of sample.

The researchers also discovered that Cas13 cleavage products can activate another Cas protein, Csm6. By including a second ssRNA construct that could be cleaved by activated Csm6 (Fig. 1b), the group boosted the signal relative to background and improved the kinetics of the SHERLOCKv2 reaction. Finally, the authors combined all of these advances into a simple assay in which a drop of sample is applied to a strip of paper that holds the diagnostic, and a colorimetric readout is given. This format requires no instruments, greatly increasing the ease with which the technology can be used by scientists and clinicians in regions of high need.

The techniques developed by Chen *et al.* and Gootenberg *et al.* both require stringent purification steps to prevent viral degradation by the body's RNA-digesting enzymes during testing. In the third paper, Myhrvold *et al.*[4] paired the previously reported version of SHERLOCK with a method for inactivating these enzymes in body fluids, allowing the authors to directly detect viruses in urine and saliva. Their approach can be used to discriminate between related viral species, which can be difficult to tell apart because they cause similar symptoms.

Viral genomes can mutate rapidly, but Myhrvold and colleagues showed that they could use their system to discriminate between sequences that differed by a single nucleotide mutation. This enabled them to identify strains of Zika virus isolated from different geographic regions, as well as a Zika virus harbouring a mutation associated with a developmental condition called microcephaly. They showed that appropriate guide sequences can be designed in less than a week, and the whole protocol implemented in less than two hours, with minimal equipment and sample processing.

The diagnostic tools described in the current papers are a major advance for the tracking and treatment of viral outbreaks. But they could also have diverse applications beyond viral detection, such as identifying tumour-specific cancer mutations for personalized medicine, and enhancing quality control in agriculture and biomanufacturing by detecting possible biological contaminants. Although their long-term shelf-life remains to be tested, the tools excel in all other ASSURED criteria set by the World Health Organization for point-of-care devices[8], and are ready for deployment. The impact and adoption of these technologies will be determined by factors such as regulatory approval, scaling up synthesis for mass production, distribution logistics and economics.

Much of the excitement about CRISPR technology has centred on its use in gene or cell therapies, which are expensive processes that, for the foreseeable future, will probably be available only in prosperous regions of the world. The current studies greatly expand and diversify the possible beneficiaries of CRISPR technology, by developing a low-cost technique that has many important potential uses. ■

D. Dewran Kocak *and* **Charles A. Gersbach**
are in the Department of Biomedical Engineering and the Center for Genomic and Computational Biology, Duke University, Durham, North Carolina 27708, USA.
e-mail: charles.gersbach@duke.edu

1. Lanciotti, R. S. *et al. Emerg. Infect. Dis.* **14**, 1232–1239 (2008).
2. Chen, J. S. *et al. Science* **360**, 436–439 (2018).
3. Gootenberg, J. S. *et al. Science* **360**, 439–444 (2018).
4. Myhrvold, C. *et al. Science* **360**, 444–448 (2018).
5. Abudayyeh, O. O. *et al. Science* **353**, aaf5573 (2016).
6. Gootenberg, J. S. *et al. Science* **356**, 438–442 (2017).
7. East-Seletsky, A. *et al. Nature* **538**, 270–273 (2016).
8. Kosack, C. S., Page, A.-L. & Klatser, P. R. *Bull. Wld Health Organ.* **95**, 639–645 (2017).

This article was published online on 30 April 2018.

NEUROSCIENCE

Models of Parkinson's disease revisited

Conventional models propose that activity levels in two neuronal pathways, which have opposing effects on movement, become imbalanced in Parkinson's disease. Analyses in mice point to a more complex reality. SEE ARTICLE P.177

THOMAS WICHMANN

Scientists have long associated structures in the forebrain called the basal ganglia with the control of movement, and with movement disorders such as Parkinson's disease. Some of the strongest evidence for the role of the basal ganglia in movement is anatomical[1]: the basal ganglia receive neuronal input from movement-related areas in the brain's frontal cortex through a region called the striatum. They then send it back to the frontal cortex. Two pathways mediate this processing, but models to explain how processing functions go awry in movement disorders are debated. On page 177, Parker *et al.*[2] add valuable data to the discussion.

In conventional models of basal-ganglia function, the two pathways that connect the striatum to output structures work in opposing ways. The first, dubbed the direct pathway, is activated by the neurotransmitter molecule dopamine, through D1-like dopamine receptors on direct-pathway spiny projection neurons (dSPNs). Activation of dSPNs is thought to facilitate movement[1,3,4]. The second, the indirect pathway, is inhibited by dopamine through D2-like receptors on indirect-pathway spiny projection neurons (iSPNs). In contrast to the first pathway, activation of iSPNs is thought to reduce movement[1,3,4].

Current models of basal-ganglia function emphasize that balanced activity in these two pathways is required for normal movement. For example, one model posits that sequential activation of the pathways could 'scale' movements — activation of dSPNs would first facilitate movements, and subsequent activation of iSPNs would terminate them[5]. Alternatively, the action-selection hypothesis proposes that interplay between dSPN and iSPN activation might assist the selection of particular actions by the frontal cortex, with initiation of certain movements associated with dSPN activation, and movement prevention associated with iSPN activation[6,7].

These models predict that movement disorders result from imbalanced pathway activities. For instance, it has conventionally been argued[1] that, in Parkinson's disease (which involves decreased striatal dopamine levels), the balance would shift in favour of iSPN activation, leading to slowness or lack of movements. By contrast, involuntary movement (dyskinesia), such as that associated with overuse of the anti-parkinsonian drug L-3,4-dihydroxyphenylalanine (L-DOPA), would arise from a shift towards dSPN activation (Fig. 1a).

Parker *et al.* used a method called fluorescence microendoscopy[8–10] to simultaneously monitor the activity and spatial arrangement of large groups of dSPNs and iSPNs in freely behaving mice. The two sets of neurons were genetically engineered to fluoresce when calcium entered the cell, indicating electrical signalling. The authors monitored three groups of mice: those under normal conditions, mice treated with a drug that causes the loss of dopamine-releasing neurons to mimic the parkinsonian state, and dopamine-depleted animals treated with L-DOPA to induce involuntary movements.

The authors' observations in normal animals

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

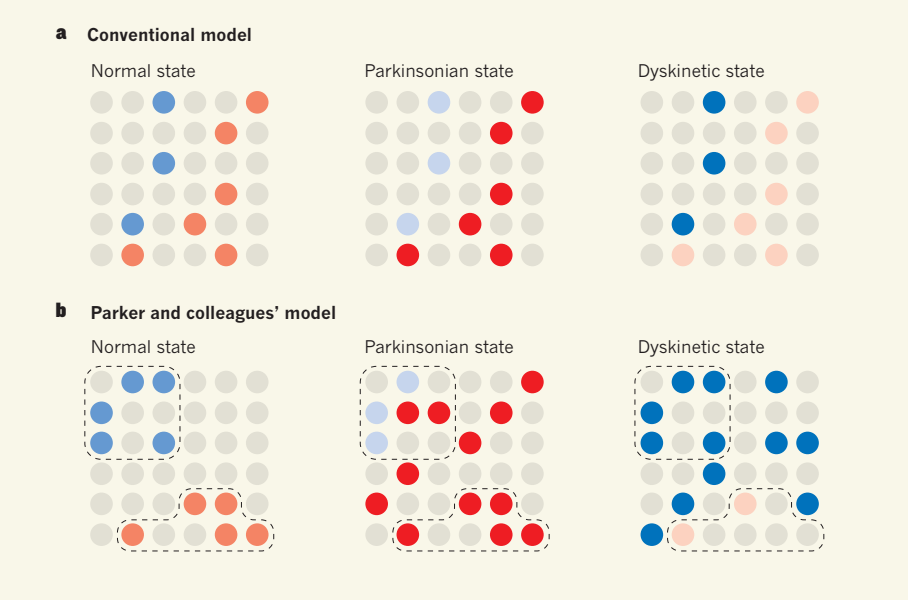


Figure 1 | Models of movement-related neuronal activation in the brain's striatum. Activation of direct-pathway spiny projection neurons (dSPNs; active neurons depicted as blue circles, with increasing colour intensity indicating increasing levels of activity) facilitates movements, whereas activation of indirect-pathway spiny projection neurons (iSPNs; red circles) inhibits movements. Grey circles indicate inactive neurons of both subtypes. **a**, Conventional models predict that normal movement is associated with activation of some dSPNs and iSPNs. In disorders such as Parkinson's disease, in which movements decrease, dSPN activation is reduced and iSPN activation is increased. The opposite changes are found in dyskinesias, which involve involuntary movements. The conventional model does not predict specific spatial arrangements of activated neurons. **b**, By contrast, Parker *et al.*[2] find that normal movement in mice involves activation of clusters (indicated in dashed lines) of dSPNs and iSPNs. In the parkinsonian state, dSPN activation is reduced and iSPN activation is increased. Clusters of activated dSPNs are smaller than in the normal state, and active iSPNs are less clustered. The opposite occurs in dyskinesias.

confirmed[9–13] that similarly sized clusters of iSPNs and dSPNs are simultaneously activated during movement. These data contradict both the scaling model (under which dSPNs are expected to be active before iSPNs) and the action-selection model (which predicts that clusters of inhibitory iSPNs should be larger than those of facilitating dSPNs). Instead, they add to mounting evidence that these pathways might not regulate the basic execution of movement, but might have a role in higher-order behavioural activities, such as the dynamic shaping of movements through reinforcement learning[14] or the control of movement 'vigour'[15].

When the authors studied dopamine-depleted parkinsonian animals, they found decreased dSPN activity and increased iSPN activity, as predicted by the conventional models. However, they also found declines over time in the movement-related activation of iSPNs (a phenomenon separate from their general increased activation level), and declines in the spatial clustering of their activation (Fig. 1b).

As expected, drugs that activate D1-like receptors increased dSPN activity in parkinsonian mice, whereas drugs that stimulated D2-like receptors reduced the elevated iSPN firing rates. L-DOPA treatment, which leads to activation of both receptors, elicited both effects and was the only treatment to reverse the clustering deficit in iSPNs, perhaps explaining its superior clinical efficacy. Finally, these beneficial effects became exaggerated in animals that developed involuntary movements induced by L-DOPA: iSPNs became underactive, and dSPNs became hyperactive and less clustered.

Parker and colleagues' study therefore supports many of the general conclusions of earlier work regarding activity changes in the normal, parkinsonian and dyskinetic states. However, it also shows that the conventional models are oversimplified, because they cannot account for the newly discovered changes in spatial clustering of active striatal neurons in disease states.

This study is an excellent example of how far experimental analyses of circuits in the basal ganglia have come in the past few years. Where we could once record a single, unidentified striatal neuron, we can now monitor hundreds of genetically characterized neurons simultaneously, albeit still at a low temporal resolution and at the expense of tissue damage caused by the implanted microendoscope. As Parker and colleagues demonstrate, these techniques can provide insights into disease mechanisms that were not possible with earlier methods. Beyond being of substantial academic interest, the insights afforded by the current study suggest that it might be worth developing therapeutic strategies for the treatment of Parkinson's disease aimed at decorrelating hyperclustered striatal activity patterns. Similar approaches have been explored using pharmacogenetic techniques[16] and deep brain stimulation[17,18] in other portions of the basal ganglia.

As with any good study, this research leads to more questions than it answers, particularly with regard to the changes in spatial clustering of active neurons seen by the authors. For instance, the mechanisms governing the opposing abnormalities in clustering in the parkinsonian and dyskinetic states are not understood, and it remains unclear whether they reflect changes in inputs to the striatum, or bona fide striatal phenomena. It is also not known whether similar clustering occurs in humans or other primates, or whether it is accompanied by altered clustering of neuron activity in other basal-ganglia structures. Only time will tell whether the gradual shifts in cluster sizes found in the parkinsonian and drug-treated animals are related to known adaptive processes, such as the loss of input-receiving structures called dendritic spines[19] on the surfaces of iSPNs and dSPNs[2], as Parker *et al.* suggest.

Finally, one of the most important questions is whether the observed changes in the spatial and temporal clustering of neuronal activities in the basal ganglia are, in fact, causally related to aspects of Parkinson's disease or the development of dyskinesias. Regardless of the answer, Parker and colleagues' study has added a new dimension to our understanding of the abnormalities seen in this disease. ■

Thomas Wichmann *is at the Yerkes National Primate Research Center, Emory University, Atlanta, Georgia 30329, USA, and in the Department of Neurology, School of Medicine, Emory University.*
e-mail: twichma@emory.edu

1. DeLong, M. R. *Trends Neurosci.* **13**, 281–285 (1990).
2. Parker, J. G. *et al. Nature* **557**, 177–182 (2018).
3. Kravitz, A. V. *et al. Nature* **466**, 622–626 (2010).
4. Freeze, B. S., Kravitz, A. V., Hammack, N., Berke, J. D. & Kreitzer, A. C. *J. Neurosci.* **33**, 18531–18539 (2013).
5. DeLong, M. & Wichmann, T. *Clin. EEG Neurosci.* **41**, 61–67 (2010).
6. Mink, J. W. *Prog. Neurobiol.* **50**, 381–425 (1996).
7. Nambu, A. *Curr. Opin. Neurobiol.* **18**, 595–604 (2008).
8. Bocarsly, M. E. *et al. Biomed. Opt. Express* **6**, 4546–4556 (2015).
9. Klaus, A. *et al. Neuron* **95**, 1171–1180 (2017).
10. Barbera, G. *et al. Neuron* **92**, 202–213 (2016).
11. Tecuapetla, F., Matias, S., Dugue, G. P., Mainen, Z. F. & Costa, R. M. *Nature Commun.* **5**, 4315 (2014).
12. Tecuapetla, F., Jin, X., Lima, S. Q. & Costa, R. M. *Cell* **166**, 703–715 (2016).
13. Cui, G. *et al. Nature* **494**, 238–242 (2013).
14. Hikosaka, O., Kim, H. F., Yasuda, M. & Yamamoto, S. *Annu. Rev. Neurosci.* **37**, 289–306 (2014).
15. Turner, R. S. & Desmurget, M. *Curr. Opin. Neurobiol.* **20**, 704–716 (2010).
16. McIver, E., Atherton, J., Surmeier, D. & Bevan, M. D. in *Neuroscience 2015* abstr. 708.711 (Soc. Neurosci., 2015).
17. Wang, J. *et al. Brain Stimul.* **9**, 609–617 (2016).
18. Adamchic, I. *et al. Mov. Disord.* **29**, 1679–1684 (2014).
19. Villalba, R. M. & Smith, Y. *J. Neural Transm.* **125**, 431–447 (2017).

This article was published online on 2 May 2018.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ECOLOGY

Precarious preferences



MICHAEL C. SINGER & CAMILLE PARMESAN

In 1993, three ecologists reported[1] that an isolated population of the butterfly *Euphydryas editha* that inhabited a meadow in Nevada was starting to evolve a preference for laying its eggs on *Plantago lanceolata* — a non-native plant introduced to the region by cattle ranchers. On page 238, two of these researchers now show[2] that the butterflies became completely dependent on the exotic plant, with adverse consequences.

Originally, *E. editha* laid its eggs on the native plant *Collinsia parviflora*, but the longer life span of *Plantago* enabled larvae to feed on it for longer, increasing larval survival. In the current study, Singer and Parmesan report that all the female butterflies they examined in 2005 preferred to lay their eggs on *Plantago* (pictured, *E. editha* resting on *Plantago*). And by 2007, all the larvae they found in the field were feeding on this non-native plant.

But cattle ranching ceased in 2005, leading to a rapid build-up of grassy vegetation and a decline in the dominance of *Plantago*. This, in turn, led to the extinction of the isolated *E. editha* population between 2007 and 2008, probably because the long grasses blocked out sunlight, cooling the warmth-loving larvae. The growth of the grasses quickly abated, but the butterfly population remained extinct until between 2013 and 2014, when *E. editha* recolonized the field, laying its eggs on *Collinsia*, and so setting the stage for the process to begin again.

These findings demonstrate how the adaptation of insect populations to human-induced environmental change can render those populations dependent on the continuation of specific human practices — a potentially precarious position, given the rapidity with which such things can change. **Anna Armstrong**

1. Singer, M. C., Thomas, C. D. & Parmesan, C. *Nature* **366**, 681–683 (1993).
2. Singer, M. C. & Parmesan, C. *Nature* **557**, 238–241 (2018).

PARTICLE PHYSICS

Weak charge of the proton measured

The proton's weak charge defines the strength of certain interactions between protons and other particles. A precise determination of this quantity provides a stringent test of the standard model of particle physics. SEE LETTER P.207

XIAOCHAO ZHENG

Subatomic particles interact through four fundamental forces. However, only two of these forces have effects on macroscopic scales: gravity keeps us grounded on Earth, and electromagnetism causes lightning on stormy days. We are not directly influenced by the other two forces — the weak and strong forces. Similarly, it is generally known that mass is at the root of gravitational interactions and that electric charges and magnetic moments are central to electromagnetism. But the physical properties that describe the strength of weak and strong interactions, known as weak and colour charges, respectively, are less familiar. On page 207, the Jefferson Lab Q_{weak} Collaboration[1] reports the first high-precision measurement of the weak charge of the proton, which sets tight constraints on physics that cannot be described by current theories.

The strong force is so overwhelming that the particles that interact through it, known as quarks and gluons, are tightly bound to one other and exist only as composite objects, such as protons and neutrons. By contrast, the weak force is so feeble that its interactions are almost completely masked by those of electromagnetism. One might therefore wonder how the weak charge of a particle can be measured if it is as small as the name implies. Fortunately, nature provides a convenient yardstick that is associated with a principle known as parity symmetry.

A process conserves parity symmetry if it occurs with the same probability as its exact mirror image. It is straightforward to see that parity symmetry is broken in the macroscopic world, particularly in biological systems. For example, most humans are right-handed. If parity symmetry were conserved for the handedness of humans, half of the population would be right-handed and half would be left-handed.

Particles also have a handedness. A right-handed particle spins in the direction defined by the curl of your four fingers when you point your right thumb along the direction of the particle's velocity. Conversely, a particle is left-handed if you must use your left hand to relate its spinning and velocity directions. Remarkably, all subatomic particles violate parity symmetry when they interact with one

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

another through the weak force. Weak charges can therefore be determined by comparing the behaviour of left- and right-handed versions of particles.

To extract the proton's weak charge, the Q_{weak} Collaboration fired beams of electrons that had a particular handedness at a proton target. They measured an asymmetry that describes the difference in the probability that right- and left-handed electrons are scattered from the proton (Fig. 1). The authors found an asymmetry of –226.5 ± 9.3 parts per billion, where the minus sign indicates that left-handed electrons are more likely to be scattered than their right-handed counterparts. To put the magnitude of this asymmetry in perspective: if parity symmetry were violated for the height of mountains, Mount Everest and its mirror-image twin would differ in height by a mere 2 millimetres, and this difference would have been measured to a precision of ± 80 μm.

The authors' result has a much higher precision than all previous experiments that studied parity violation by scattering electrons from a nuclear target. The E158 experiment at the SLAC National Accelerator Laboratory in Menlo Park, California, had a comparable precision, but measured the weak charge of the electron rather than that of the proton[2]. The Q_{weak} Collaboration used its measured asymmetry to determine that the proton's weak charge is 0.0719 ± 0.0045, which is in excellent agreement with the value predicted by the standard model of particle physics[3]. For comparison, in the convention used by the authors, the proton's electric charge is +1.

One might question why physicists want to measure the proton's weak charge to such high precision. The short answer: to test the limits of our knowledge. At a basic level, physicists seek to discover if, and at what length scale, current theories fail to explain observational data. Such a failure could imply the existence of a fifth fundamental force — a previously undiscovered type of interaction that has a role at energies higher than have been explored so far.

The measurement reported by the Q_{weak} Collaboration shows that such interactions, if they exist, would reveal themselves at particle energies beyond several TeV (1 TeV is 10^{12} electronvolts). For comparison, the energies released in nuclear-fission reactors, in which nuclei split into two or more fragments, are at the level of 10^6 eV per particle. The authors' lower limit for the energy scale of new physics is comparable to, and complements, that set by experiments at the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva, Switzerland[4,5]. This is remarkable, given that the energy of the authors' electron beams is thousands of times lower than that of the Large Hadron Collider's proton beams.

More than a century ago it was demonstrated that electric charge comes in discrete chunks[6], which provided a bridge between classical electromagnetism and modern quantum mechanics. By measuring quantities — from physical constants to particle properties such as the proton's weak charge — at increasingly higher precision, new interactions and particles could be discovered that require current theories to be revised. Such lofty pursuits can be maintained only if research is carried out by multiple groups across generations, as opposed to single groups in isolated instances. One day, the knowledge acquired might lead to a breakthrough. In the meantime, it will make our world a better one by providing a deeper understanding of the physical laws of nature. ■

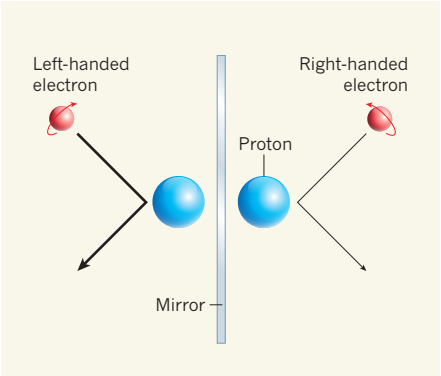


Figure 1 | Parity violation in electron–proton scattering. Electrons can be either left- or right-handed, depending on the direction in which they spin (red arrows) relative to their motion (black arrows). The Jefferson Lab Q_{weak} Collaboration[1] reports that left-handed electrons are slightly more likely to be scattered from a proton target than are their right-handed counterparts. This result violates a principle known as parity symmetry because the two scattering processes are mirror images of each other. The authors used the measured asymmetry to determine the value of a fundamental physical quantity known as the weak charge of the proton to an unprecedented precision.

Xiaochao Zheng *is in the Department of Physics, University of Virginia, Charlottesville, Virginia 22904, USA.*
e-mail: xz5y@virginia.edu

1. The Jefferson Lab Q_{weak} Collaboration. *Nature* **557**, 207–211 (2018).
2. SLAC E158 Collaboration. *Phys. Rev. Lett.* **95**, 081601 (2005).
3. Patrignani, C. *et al.* (Particle Data Group). *Chin. Phys. C* **40**, 100001 (2016).
4. ATLAS Collaboration. *Eur. Phys. J. C* **74**, 3134 (2014).
5. CMS Collaboration. *Phys. Lett. B* **746**, 79–99 (2015).
6. Millikan, R. A. *Phys. Rev.* **2**, 109–143 (1913).

NEUROSCIENCE

Connections that control defence strategy

When presented with a threat, most mice freeze or hide, but a few respond more aggressively. The brain circuits underlying this behavioural choice have now been unpicked. SEE ARTICLE P.183

DAYU LIN

My nine-year-old daughter got a pair of mice for Christmas. To acclimatize her new pets to her presence, she put her hands into their cage. One of the mice ran straight to its shelter and hid; the other stayed in a corner, held its tail up against the glass wall and rattled it rapidly. Tail rattling is a common response to stress across species (most famously in snakes[1]), and is considered a warning to intruders[1,2]. But why do some mice hide, and others threaten? What brain mechanisms determine which defensive strategy to deploy in the face of potential danger? On page 183, Salay *et al.*[3] reveal a brain region that could be responsible for this decision in mice.

Salay *et al.* first surveyed the brain regions that are activated by threatening cues. They placed mice in a test arena, then presented them with an overhead looming stimulus — an expanding dark-coloured circle, to mimic an approaching predator. When the mice detected the looming cues, they either froze or quickly ran towards a shelter under which they could hide from the stimulus. Occasionally, mice rattled their tails once they had reached the safety of the shelter.

To identify the neurons that determine these behavioural responses, the researchers analysed the protein c-Fos, which is expressed rapidly in neurons after they have been highly active. Many brain regions show consistently high levels of c-Fos in response to looming, and one caught the authors' attention — the ventral midline thalamus (vMT). The vMT is interesting in that it is not a part of the eye-to-brain visual pathway or a motor pathway, as might be expected for neurons involved in processing visual stimuli such as the looming cues in this study. Instead, it receives diffuse inputs from limbic areas and the midbrain (regions that support emotion, motivational behaviours and bodily responses), and projects heavily to higher cognitive areas, such as the prefrontal cortex[4]. Thus, this region is well suited to signalling the internal state of the animal and guiding its defence strategies.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

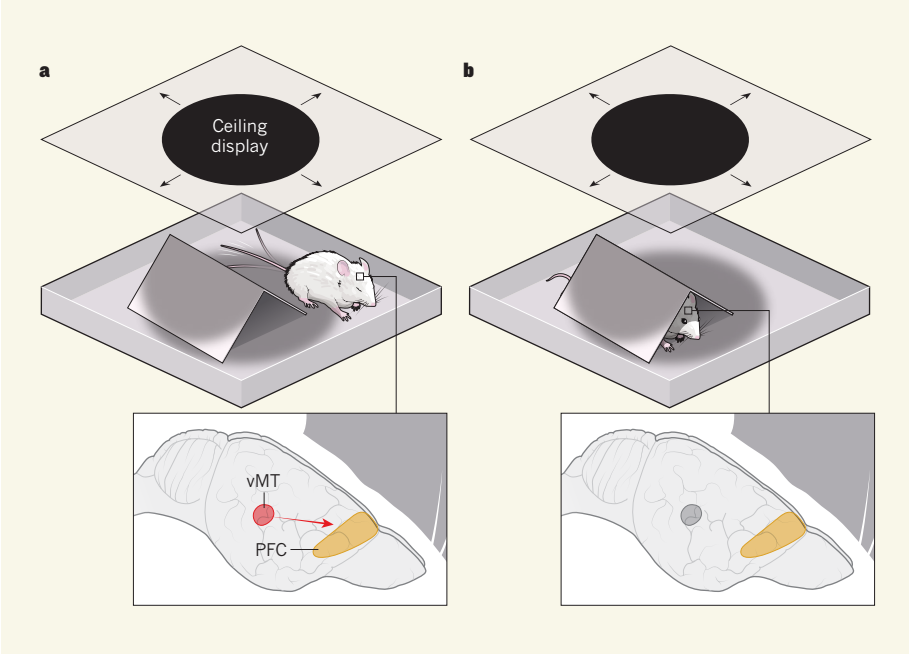


Figure 1 | To recede or to challenge. Salay *et al.*[3] report that a region of the brain known as the ventral midline thalamus (vMT) controls how a mouse responds to a threatening visual stimulus, such as an expanding dark circle in a ceiling display that looms overhead. **a**, The authors provide evidence that signalling from the vMT to the prefrontal cortex (PFC) promotes an active defence strategy, such as making an aggressive tail-rattling motion. **b**, By contrast, when levels of vMT activation are lower, mice respond to the looming stimulus by hiding.

In vivo recordings of neuronal activity revealed that vMT cells are most active when the animals are responding actively to looming stimuli by running and tail rattling. Salay *et al.* therefore asked whether vMT activity is sufficient to guide an animal's defence strategy. The authors injected the vMT with a viral construct carrying a receptor protein that, when activated, excites neurons. The receptor is activated on binding to a small ligand, which the authors administered systemically weeks after injection of the virus. Strikingly, when Salay and colleagues artificially activated vMT cells, the mice no longer hid from the looming stimuli, but instead rattled their tails frequently and ran around the open space. Conversely, when vMT cells were suppressed using a similar approach, the animals spent more time freezing and never rattled their tails in response to looming stimuli.

To investigate how vMT activation affects responses to looming in more detail, the authors genetically modified cells such that they could be activated by a laser, giving millisecond-level control of cell activity. They found that activating vMT cells either in tandem with or 30 seconds before the presentation of looming stimuli promoted the same active defence behaviours, such as tail rattling. This behavioural shift could outlast the stimulation itself, which makes it likely that activating the vMT causes a shift in internal state, instead of eliciting acute motor actions. Consistent with this hypothesis, the researchers showed that vMT activation induces strong and long-lasting autonomic responses indicative of increased arousal, such as pupil dilation, but does not induce changes in motor activity in the absence of looming stimuli.

Next, Salay *et al.* investigated which projections from the vMT are relevant for guiding defence behaviours. They found that distinct regions of the vMT project to the prefrontal cortex and basolateral amygdala, both of which are implicated in many complex functions, including fear and anxiety. The authors again used a viral technique, this time to manipulate specific vMT outputs, and found that activating vMT projections to the prefrontal cortex pathway promoted an active threat-coping strategy (tail rattling) and increased arousal (Fig. 1). By contrast, activating the vMT–basolateral amygdala projection favoured a different behaviour, freezing. Taken together, the authors conclude that distinct vMT outputs control opposing threat-coping strategies.

Escaping and hiding are basic survival instincts when an animal is faced with a powerful predator. However, from time to time, prey animals need to stand their ground. For instance, adult ground squirrels protect pups in a burrow fiercely — after detecting a nearby snake, the squirrel flags its tail and aggressively harasses the snake by biting and kicking dirt towards it. As a result, the snake typically abandons its hunting effort and moves away[5]. When a weak animal challenges a stronger predator in this way, the animal might be considered 'brave'. Does it stand to reason, then, that the vMT drives a 'courageous' state?

Answers to this question remain unclear,



50 Years Ago

Britain's first heart transplant operation, performed last week at the National Heart Hospital, has been received with the usual gush of publicity which accompanies such events. But less acclaim has been given to the liver transplantation operation carried out at Addenbrooke's Hospital, Cambridge, on May 2, although this is technically by far the trickier of the two operations … Several attempts at liver transplants have been made in the United States, but met with little success until June of last year. Since then advances in immunosuppression and organ storage have enabled the livers of deceased persons to be maintained in a form acceptable to the recipient's body for several hours … The heart surgery team at the National Heart Hospital says that the heart transplant operation is the first of a series of operations leading to a combined heart–lung transplant before the end of the year.

From *Nature* 11 May 1968

100 Years Ago

The Gas Traction Committee, appointed in November last to consider the employment of coal-gas as a substitute for petrol and petroleum products in motors, its general safety, and conditions for use, has issued an interim report … This deals chiefly with the present use of gas, mainly at low pressures, in suitable fabric bags; the work is to be continued to cover its use under higher compressions when the necessary appliances and plant are obtainable. The Committee is satisfied that gas can be efficiently, safely, and promptly substituted for motor spirit (only two minor accidents have been reported). Two hundred and fifty cubic feet are considered equivalent to one gallon of petrol, so that gas at 4*s*. per 1000 cub. ft. is equivalent in cost to petrol at 1*s*. per gallon.

From *Nature* 9 May 1918

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

because a direct assessment of courage is not possible in non-human species. Intriguingly, however, the internal state elicited by vMT activation seems to be perceived as positive, because the mice prefer to stay in a chamber in which they can receive the vMT stimulation, given a choice. Future studies that examine the correlation between vMT activity and courage in humans will help us to further understand the emotional state encoded by the vMT.

It is also still unclear whether tail rattling in mice can effectively deter a predator in the wild. But this behaviour did successfully stop my daughter from bothering her pets. ■

Dayu Lin *is at the Neuroscience Institute, New York University School of Medicine, New York, New York 10016, USA.*
e-mail: dayu.lin@nyumc.org

1. Greene, H. W. *J. Herpetol.* **7,** 143–161 (1973).
2. Haber, S. B. & Simmel, E. C. *Bull. Psychonom. Soc.* **7,** 84–86 (1976).
3. Salay, L. D., Ishiko, N. & Huberman, A. D. *Nature* **557**, 183–189 (2018).
4. Van der Werf, Y. D., Witter, M. P. & Groenewegen, H. J. *Brain Res. Rev.* **39,** 107–140 (2002).
5. Clark, R. W. *Behav. Ecol. Sociobiol.* **59,** 258–261 (2005).

This article was published online on 2 May 2018.

STRUCTURAL BIOLOGY

A closer look at human telomerase

The telomerase enzyme performs a crucial role by maintaining protective caps of repetitive DNA sequences at the ends of chromosomes. A structure of human telomerase casts fresh light on its architecture. SEE ARTICLE P.190

MICHAEL D. STONE

The chromosomes of eukaryotic organisms, which include plants, animals and fungi, are capped by protective structures called telomeres[1]. In certain types of proliferative cell, these structures are maintained by the telomerase enzyme. Inherited genetic mutations that compromise telomerase function cause disorders characterized by deterioration of proliferative tissues[2], whereas increased expression of telomerase supports unbridled cell growth in most human cancers[3]. Efforts to develop telomerase-based therapeutics have been hampered by an incomplete understanding of the structure and organization of the telomerase complex. On page 190, Nguyen *et al.*[4] report a structure of human telomerase bound to DNA, providing an unprecedented view of how the enzyme complex is organized and establishing a framework for drug discovery.

Telomerase is a ribonucleoprotein (RNP) enzyme — part RNA, part protein. To retain normal function, the enzyme must contain at least a telomerase reverse transcriptase (TERT) protein subunit and a long, non-coding telomerase RNA (TR). Telomerase complexes also usually include several species-specific proteins that are important for regulating RNP assembly[5]. The enzyme maintains telomeres by catalysing the addition of simple DNA-sequence repeats onto the ends of linear chromosomes. The catalytic TERT subunit uses a small portion of TR as a template for these repeats.

Vertebrate telomerases also belong to the family of box H/ACA RNP complexes, which generally modify essential cellular RNAs such as transfer RNAs, ribosomal RNA and spliceosomal RNA[6]. Curiously, despite possessing the RNA and protein components of H/ACA RNPs, telomerase does not exhibit the RNA-modifying activity typically associated with this class of complex. Nevertheless, mutations in those components impair telomere maintenance and cause diseases related to telomerase deficiency[2].

Human telomerase has low abundance in cells, and attempts to reconstitute it from its purified components have been thwarted by protein insolubility and the low efficiency of RNP assembly. This challenge can be overcome by expressing TERT and TR in cultured human cells that naturally express large quantities of the other telomerase proteins needed for RNP assembly[7]. This approach was used to reconstitute human telomerase and produce the first low-resolution structure[8] of the enzyme using electron microscopy (EM). However, telomerase complexes made in this way can contain different subunits and have varying structures, both of which limit the resolution of EM structures.

Nguyen *et al.* report an improved purification procedure that yields substantially more-active and more homogeneously assembled telomerase. Armed with this highly purified enzyme, the authors used cryo-electron microscopy (cryoEM) to reconstruct the structure of human telomerase at subnanometre resolution. This resolution is three times better than that of the previously reported structure, and allowed the authors to reliably work out how the various telomerase components

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

correlate with the regions of electron density mapped out by cryoEM.

The authors' structure has a bilobal shape (Fig. 1) consistent with the earlier low-resolution structure. This shape was previously interpreted[8] as a telomerase dimer — that is, each lobe was thought to contain a copy of the TERT and TR subunits. Nguyen *et al.* now propose a markedly different interpretation. They point out that the structure of one of the lobes is consistent with high-resolution models[9] of TERT and TR, which comprise the 'catalytic core' of the telomerase complex. They also compare their data with the crystal structure[10] of an H/ACA RNP bound to a single 'hairpin' RNA, and postulate that the second lobe contains the H/ACA RNP. Moreover, they propose that this H/ACA lobe includes a single copy of the TCAB1 protein, which regulates telomerase trafficking in the nucleus[11]. The bilobal architecture results from a remarkably extended RNA scaffold, consistent with the general model that TRs provide a flexible scaffold for telomerase-associated proteins[12]. The two distinct lobes interact through two bridging RNA helices.

The discrepancies between Nguyen and colleagues' structure and previous reports of the composition and organization of human telomerase are probably rooted in the different methods used to purify the enzyme, which might yield distinct telomerase subcomplexes and, in some cases, promote formation of a telomerase dimer. Nevertheless, the present structure supports the hypothesis that telomerase from a range of organisms can function as a monomer[13,14].

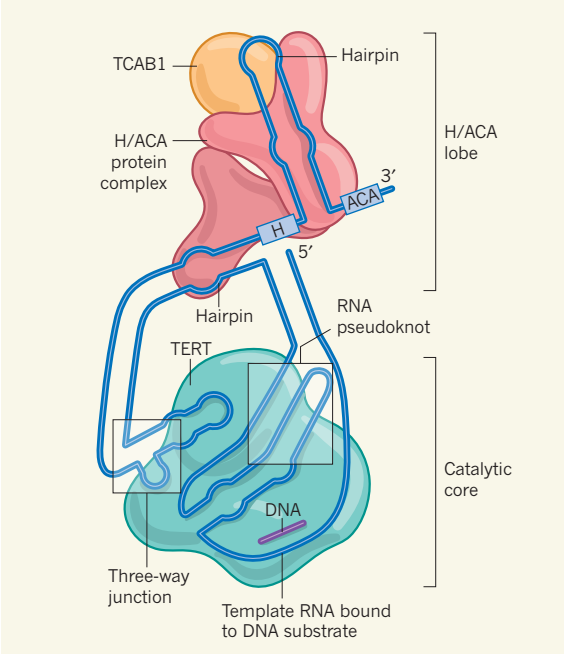


Figure 1 | Cartoon of the human telomerase enzyme. Telomerase enzymes catalyse the addition of short DNA-sequence repeats to protective caps (telomeres) at the ends of linear chromosomes. Nguyen *et al.*[4] report a structure of human telomerase obtained using cryo-electron microscopy. The overall structure is bilobal, and consists of an extended RNA (blue) that acts as a scaffold for various proteins. The authors suggest that the H/ACA lobe — named after the H and ACA motifs within the RNA — contains two copies of the tetrameric H/ACA protein complex bound to two 'hairpin' regions of the RNA scaffold, as well as a TCAB1 protein. This lobe facilitates the assembly of the telomerase enzyme from its components, and controls trafficking of the enzyme in the nucleus. The other lobe is proposed to contain the enzyme's catalytic core, which includes the TERT protein, together with the pseudoknot and three-way junction of the RNA. The RNA sequence that is used as a template for the DNA repeats is also in the catalytic core, bound to a DNA substrate.

In Nguyen and co-workers' reconstruction of the catalytic core of human telomerase, the active site of the ring-shaped TERT subunit contains the template RNA bound to a DNA substrate. The two regions of TR required for catalytic function *in vitro* are known as the template/pseudoknot domain (t/PK) and the conserved regions 4/5 (CR4/5) domain. In the authors' structure, the evolutionarily conserved pseudoknot fold in the t/PK domain resides on the 'back' side of the TERT ring, far from the active site. This is highly consistent with the pseudoknot's location in the previously reported structure[13] of telomerase from the protozoan *Tetrahymena thermophila*, and suggests that this essential TR element indirectly promotes telomerase function, perhaps by guiding proper RNP assembly or by influencing the conformation of TERT subdomains during catalysis.

The authors' model of the H/ACA lobe is the first structure of a eukaryotic H/ACA RNP, and includes two complete sets of the four H/ACA proteins (dyskerin, NOP10, NHP2 and GAR1) bound to RNA hairpins. The structure reveals interesting differences in the way in which each set of H/ACA proteins interacts with its respective hairpin. The foundation of the H/ACA lobe structure is made by dyskerin subunits, which bind to the base of each RNA hairpin and also make extensive contacts with each other. The remaining three H/ACA proteins seem to make numerous contacts with one of the RNA hairpins, but substantially fewer with the other. Lastly, the TCAB1 protein binds to a structure within TR known as the CAB box, found at the apex of the H/ACA lobe. These results provide a structural explanation for a wealth of previously reported biochemical data[15,16] that defined the subunit stoichiometry, as well as the RNA-hairpin sequence and structure requirements for stable assembly, of the TR H/ACA RNP.

The structural segregation of the catalytic core and H/ACA lobe raises the question of why the telomerase complex has retained the large H/ACA RNP throughout evolution. Part of the answer is likely to be that TCAB1 is needed to mediate trafficking of telomerase components during RNP assembly. However, another striking feature of the H/ACA lobe is that its proteins seem to promote a specific conformation of TR that brings the CR4/5 domain close to the catalytic core, where it is needed to support catalysis. This bears a remarkable resemblance to the *T. thermophila* telomerase RNP, in which an RNA-binding protein called p65 is required to remodel TR and promote contacts with the catalytic TERT subunit[13,17]. This resemblance suggests that telomerases from different organisms have co-opted different types of RNA-binding protein to promote RNP assembly. The idea that the H/ACA proteins direct RNA folding can be tested experimentally, using the new structural model as a framework with which to design and interpret incisive experiments.

Nguyen and colleagues' work sets a new bar for structural models of human telomerase, but there is still work to be done. A complete view of telomerase structure and function will require atomic-resolution models of the enzyme in its various functional states. Advances in cryoEM techniques, together with other structural and biophysical tools, will no doubt continue to enhance our appreciation of the intricate workings of this fascinating enzyme. ■

Michael D. Stone *is in the Chemistry and Biochemistry Department and at the Center for Molecular Biology of RNA, University of California, Santa Cruz, Santa Cruz, California 95064, USA.*
e-mail: mds@ucsc.edu

1. Palm, W. & de Lange, T. *Annu. Rev. Genet.* **42**, 301–334 (2008).
2. Vulliamy, T. J. & Dokal, I. *Biochimie* **90**, 122–130 (2008).
3. Kim, N. W. *et al. Science* **266**, 2011–2015 (1994).
4. Nguyen, T. H. D. *et al. Nature* **557**, 190–195 (2018).
5. Egan, E. D. & Collins, K. *RNA* **18**, 1747–1759 (2012).
6. Kiss, T., Fayet-Lebaron, E. & Jády, B. E. *Mol. Cell* **37**, 597–606 (2010).
7. Cristofari, G. & Lingner, J. *EMBO J.* **25**, 565–574 (2006).
8. Sauerwald, A. *et al. Nature Struct. Mol. Biol.* **20**, 454–460 (2013).
9. Chan, H., Wang, Y. & Feigon, J. *Annu. Rev. Biophys.* **46**, 199–225 (2017).
10. Li, L. & Ye, K. *Nature* **443**, 302–307 (2006).
11. Venteicher, A. S. *et al. Science* **323**, 644–648 (2009).
12. Zappulla, D. C. & Cech, T. R. *Proc. Natl Acad. Sci. USA* **101**, 10024–10029 (2004).
13. Jiang, J. *et al. Nature* **496**, 187–192 (2013).
14. Alves, D. *et al. Nature Chem. Biol.* **4**, 287–289 (2008).
15. Egan, E. D. & Collins, K. *Mol. Cell. Biol.* **30**, 2775–2786 (2010).
16. Egan, E. D. & Collins, K. *Mol. Cell. Biol.* **32**, 2428–2439 (2012).
17. Stone, M. D. *et al. Nature* **446**, 458–461 (2007).

This article was published online on 25 April 2018.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0090-6

Diametric neural ensemble dynamics in parkinsonian and dyskinetic states

Jones G. Parker[1,2,9], Jesse D. Marshall[1,3,5,9], Biafra Ahanonu[1,3], Yu-Wei Wu[4], Tony Hyun Kim[1], Benjamin F. Grewe[1,6], Yanping Zhang[1,3], Jin Zhong Li[1,7], Jun B. Ding[4], Michael D. Ehlers[2,8]* & Mark J. Schnitzer[1,3]*

Loss of dopamine in Parkinson's disease is hypothesized to impede movement by inducing hypo- and hyperactivity in striatal spiny projection neurons (SPNs) of the direct (dSPNs) and indirect (iSPNs) pathways in the basal ganglia, respectively. The opposite imbalance might underlie hyperkinetic abnormalities, such as dyskinesia caused by treatment of Parkinson's disease with the dopamine precursor L-DOPA. Here we monitored thousands of SPNs in behaving mice, before and after dopamine depletion and during L-DOPA-induced dyskinesia. Normally, intermingled clusters of dSPNs and iSPNs coactivated before movement. Dopamine depletion unbalanced SPN activity rates and disrupted the movement-encoding iSPN clusters. Matching their clinical efficacy, L-DOPA or agonism of the D_2 dopamine receptor reversed these abnormalities more effectively than agonism of the D_1 dopamine receptor. The opposite pathophysiology arose in L-DOPA-induced dyskinesia, during which iSPNs showed hypoactivity and dSPNs showed unclustered hyperactivity. Therefore, both the spatiotemporal profiles and rates of SPN activity appear crucial to striatal function, and next-generation treatments for basal ganglia disorders should target both facets of striatal activity.

According to the classical rate model of Parkinson's disease, dSPNs and iSPNs normally have balanced activity rates, and the loss of dopamine induces an activity imbalance that suppresses movement by enhancing basal ganglia inhibition of the thalamo-cortical motor system[1–4]. The premise of the rate model, that the direct and indirect pathways are functionally opposed, concisely explained multiple aspects of basal ganglia function and dysfunction and led to descriptions of dystonia and Huntington's disease[1,2], Tourette's syndrome[5], dyskinesia[6], addiction[7], chronic pain[8], depression[9] and schizophrenia[10] based on imbalanced SPN activity rates. However, without cellular-level recordings that enable the comparison of dSPN and iSPN activity levels, no definitive evidence of a striatal activity imbalance has been shown in any of these conditions.

Optogenetic studies support the idea that dSPNs and iSPNs have opposing roles in movement initiation[11]. However, fluorescence recordings have recently shown that both cell types normally activate at movement onset with similar time courses[12–14]. These results challenge the rate model and reinvigorate models in which the direct and indirect pathways coordinate action selection by concurrently activating and suppressing competing motor programs, respectively[15]. The two pathways could still be opposed in some manner, but the oversimplicity of the rate model may underlie its inability to explain several notable features of Parkinson's disease[16], including the loss of movement-specific activity in nuclei downstream of the striatum[17]. Therefore, imbalanced SPN dynamics might only partially account for the motor symptoms of Parkinson's disease.

L-DOPA is the mainstay therapy for treating these symptoms, but whether it alters SPN activity and why it is superior to dopamine-receptor agonism remain unknown[18]. Notably, the efficacy of L-DOPA gradually decreases, as patients increasingly exhibit involuntary movements termed L-DOPA-induced dyskinesia (LID)[18]. Effective alternatives are needed to delay L-DOPA treatment and supplement or replace it upon loss of efficacy. LID is not fully understood, but might involve an activity imbalance opposite to that found in Parkinson's disease, with dSPN hyperactivity[19]. However, as with Parkinson's disease direct proof for these ideas has been lacking. To examine these issues, we compared dSPN and iSPN dynamics in freely behaving mice, under normal and parkinsonian conditions, before and after common treatments for Parkinson's disease and during LID.

Striatal encoding of movement

To selectively monitor dSPNs or iSPNs in the dorsomedial striatum, we respectively injected *Drd1a*^{cre} (*Drd1a* is also known as *Drd1*) or *Adora2a*^{cre} driver mice with a virus mediating Cre-dependent expression of the green-fluorescent Ca^{2+}-indicator GCaMP6m[20,21] (Methods). Control studies in live striatal tissue slices confirmed that somatic Ca^{2+} dynamics report spiking equivalently in dSPNs and iSPNs (Extended Data Fig. 1a–i).

To track SPN activity over weeks without affecting locomotor behaviour, we used a head-mounted microscope[22,23] and time-lapse microendoscopy[24] to image SPN dynamics as mice explored an open arena (Fig. 1 and Extended Data Fig. 1j, k). We computationally extracted activity traces of individual SPNs from the Ca^{2+} movies (262 ± 11 dSPNs (mean ± s.e.m.), 277 ± 8 iSPNs per mouse per session; 369 total 1-h-long sessions in 17 *Drd1a*^{cre} mice, 21 *Adora2a*^{cre} mice). Ca^{2+} transient events occurred at similar rates for dSPNs and iSPNs and had waveforms consistent with those in striatal slices (Extended Data Fig. 1l–n).

During movement, Ca^{2+} event rates rose markedly in both SPN types with similar dependencies on locomotor speed (Fig. 2a, b and Supplementary Videos 1, 2). Averaged over all Ca^{2+} events, locomotor speed rose after Ca^{2+} excitation with indistinguishable latencies in dSPNs and iSPNs (192 ± 9 ms (mean ± s.e.m.) and 196 ± 11 ms, respectively; $P = 0.8$; rank-sum test). Averaged over all movement bouts, mean SPN activity rose approximately 1 s before motion onset (Fig. 2c), with mean times to reach the half-maximum amplitude that were indistinguishable given the approximately 40-ms resolution of our statistical analysis (*Drd1a*^{cre} mice, -0.53 ± 0.04 s (mean ± s.e.m.)

[1]CNC Program, Stanford University, Stanford, CA, USA. [2]Neuroscience Research Unit, Pfizer Inc., Cambridge, MA, USA. [3]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. [4]Department of Neurosurgery, Stanford University School of Medicine, Stanford, CA, USA. [5]Present address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [6]Present address: Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland. [7]Present address: Cegeim Bio-Engineering (Changchun) Co. Ltd., Jilin, China. [8]Present address: Biogen, Cambridge, MA, USA. [9]These authors contributed equally: Jones G. Parker, Jesse D. Marshall. *e-mail: mschnitz@stanford.edu; michael.ehlers@biogen.com

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

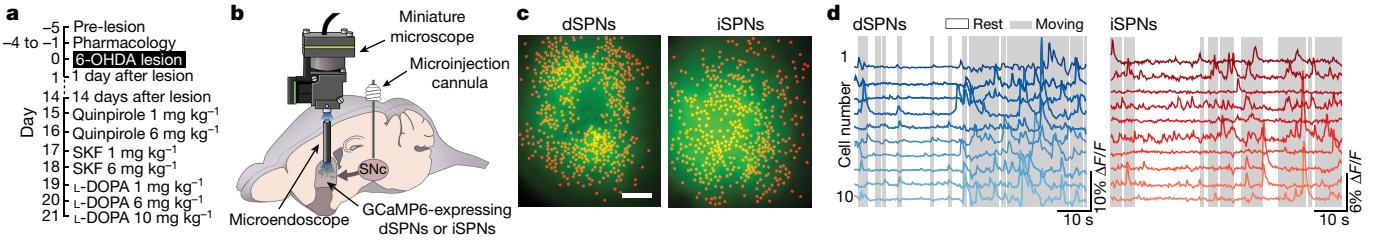


Fig. 1 | Ca^{2+} imaging of SPNs in freely behaving mice. a, Timeline of Ca^{2+} imaging sessions, 6-OHDA infusion and drug treatments. SKF, SKF81297. **b**, An Inscopix miniature microscope and microendoscope imaged SPN activity patterns. Infusion of 6-OHDA into substantia nigra (SNc) on day 0 ablated dopamine cells. **c**, Somata (red) of 573 dSPNs and 356 iSPNs in example *Drd1a*^{cre} (left) and *Adora2a*^{cre} (right) mice, atop mean fluorescence images of the dorsomedial striatum. Scale bar, 125 μm. **d**, Example Ca^{2+} activity traces in mice exploring an open arena. Grey shading, running periods; white shading, resting periods.

relative to motion onset, 17 mice, 5,372 cells, 2,899 movement bouts; *Adora2a*^{cre} mice, −0.48 ± 0.04 s, 21 mice, 5,880 cells, 3,559 movement bouts; $P = 0.7$; rank-sum test).

SPN activity was markedly clustered during movement, although the clusters of coactive cells were not absolutely stereotyped[14]. The spatial coordination of coactive SPNs rose and fell at motion onset and offset, respectively, with time courses that resembled those of the Ca^{2+} activity (Fig. 2c and Extended Data Fig. 2). Cell pairs within 100 μm or less were especially coactive, but beyond this distance pairwise activity correlations gradually decreased (Fig. 2d, e). By comparing the sets of cells that were activated during five different types of movement, we found that SPNs activated during bouts of the same movement type were, on average, more similar and closer together than those activated during different movement types[13] (Extended Data Fig. 3 and Methods).

To examine how dSPN and iSPN activity patterns were inter-related, we imaged their concurrent dynamics by two-photon microscopy in head-fixed mice in which both SPN-types expressed GCaMP6m and only dSPNs expressed the red-fluorescent marker tdTomato (Extended Data Fig. 4 and Supplementary Video 3). As the mice ran in place, dSPN and iSPN dynamics tracked locomotor behaviour indistinguishably. Nearby (20–100 μm) SPNs were equally coactive with other SPNs regardless of type, and the two cell populations encoded running speed equivalently.

Striatal activity after dopamine depletion

We next examined SPN dynamics in a classic model of Parkinson's disease[25]. Ipsilateral to the imaged striatum, we lesioned dopamine cells of the subtantia nigra pars compacta (SNc) using the neurotoxin 6-OHDA. Immunostaining for tyrosine hydroxylase, a marker of dopamine neurons, confirmed that the lesioned SNc lacked dopamine cell bodies 24 h after 6-OHDA infusion (Extended Data Fig. 5a, b). Following Ca^{2+} imaging, immunofluorescence intensities in the ipsilateral striatum more than 17 days after 6-OHDA infusion were 6 ± 1% (mean ± s.e.m.) of those contralateral to the lesion ($P = 6 \times 10^{-8}$; $n = 25$ mice; signed-rank test; Fig. 3a), verifying loss of dopamine neuron axons.

One day after the mice received 6-OHDA, locomotion was reduced relative to mice that had received saline infusions (Fig. 3b, c and Extended Data Fig. 5c). During Ca^{2+} imaging, background fluorescence was also reduced, but with almost no effect on the detection of Ca^{2+} events or SPNs (Extended Data Fig. 5d–h). Consistent with the rate model of Parkinson's disease, post-lesion (1 day) activity rates during rest and locomotion were decreased in dSPNs and increased in iSPNs compared to pre-lesion values (Fig. 3d–f and Extended Data Fig. 6). These effects resembled those of acute treatment with either D_1 dopamine receptor (D1R) or D_2 dopamine receptor (D2R) antagonists (Extended Data Fig. 7), which can cause parkinsonism in humans[26].

By 14 days after the lesion, a time point that may be more representative of the sustained dopamine depletion in patients with Parkinson's disease[27], locomotor activity had risen but remained below normal levels (Fig. 3c). iSPN activity remained elevated at baseline but barely rose at movement onset, no longer reflected locomotor speed and had lost nearly all of its spatial coordination (Fig. 3d–j). These abnormalities were also present during a rotarod assay of motor coordination (Extended Data Fig. 6c–e). The iSPN ensembles activated during different types of movement also became less distinctive after dopamine depletion, indicating reduced encoding specificity (Extended Data Fig. 8). By comparison, dSPN activity rates remained consistently depressed but still rose at motion onset and retained spatial coordination, a clear relationship to locomotion and selectivity for specific movement types (Fig. 3d–j). Thus, between acute and extended dopamine depletion, iSPN but not dSPN ensembles underwent pronounced changes in how their spatiotemporal dynamics encoded locomotion.

Treatment of aberrant neural dynamics

We next examined whether dopamine replacement or receptor agonism could reverse these aberrant SPN dynamics. Across days 15–20 after

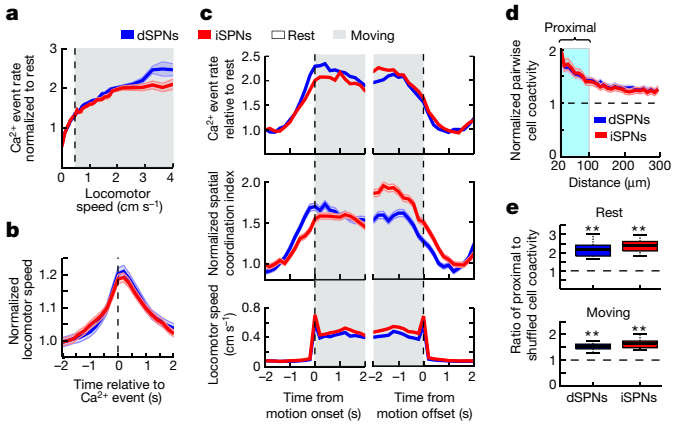


Fig. 2 | SPNs encode movement via spatially clustered bursts of activity. **a**, Mean SPN Ca^{2+} event rates versus running speed, relative to mean rates in resting mice (<0.5 cm s^{-1}). Grey shading denotes locomotion (>0.5 cm s^{-1}) in **a** and **c**. **b**, Mean speed relative to occurrences of Ca^{2+} events, normalized to mean speed 2 s before Ca^{2+} excitation. Traces are averages across all dSPNs (blue trace) and iSPNs (red trace). **c**, Mean speed (bottom), Ca^{2+} event rate (top) and spatial coordination of SPN activity (middle), relative to times of motion onset and offset. Top two traces are normalized to mean values in resting mice. Kinks in speed traces at motion onset/offset reflect the definition of a motion bout, for which the minimum speed to initiate a bout exceeds that to sustain one. **d**, Mean pairwise coactivity during locomotion versus the separation of cell pairs, normalized to values in temporally shuffled datasets (dashed line). Cyan, proximal (20–100 μm) cell pairs. **e**, Coactivity of proximal cell pairs exceeded that of the activity in shuffled datasets ($^{**}P < 5 \times 10^{-4}$; signed-rank test comparing real and shuffled data; $n = 17$ *Drd1a*^{cre}, 21 *Adora2a*^{cre} mice). Exact P values can be found in the Supplementary Information for all figures. **a**–**e**, Data based on the same recordings (day −5; 17 *Drd1a*^{cre}, 21 *Adora2a*^{cre} mice). **a**–**d**, Colour shading indicates the s.e.m. **e**, In box-and-whisker plots, the horizontal lines denote median values, boxes cover the middle two quartiles and whiskers span 1.5× the interquartile range. Box-and-whisker plots in all other figures are formatted identically.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

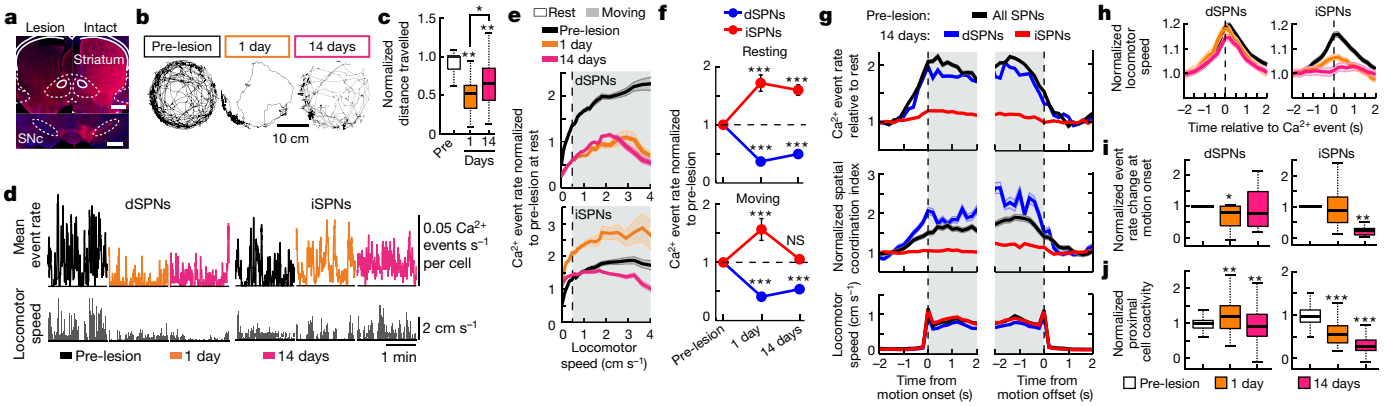


Fig. 3 | Dopamine loss differentially alters dSPN and iSPN activity patterns. a, Brain slices were immunostained for tyrosine hydroxylase (red) >14 days after unilateral 6-OHDA lesion and show unilateral loss of dopamine cell bodies and axons. White lines, brain area boundaries[38]. DAPI was used to stain the nuclei (blue). Scale bars, 1 mm. **b**, Example locomotor trajectories (15 min), before 6-OHDA lesion and 1 and 14 days after the lesion. **c**, Distance travelled, normalized for each mouse to its pre-lesion value. $^{**}P < 3 \times 10^{-3}$ compared to pre-lesion; $^{*}P = 0.013$ comparing 1 day to >14 days; signed-rank test, $n = 25$ mice. **d**, Traces of locomotor speed and mean Ca^{2+} event rate in example *Drd1a*^{cre} (left) and *Adora2a*^{cre} (right) mice, before (black), 1 day (orange) and 14 days (pink) after 6-OHDA lesion. **e**, Mean Ca^{2+} event rates versus speed, before and after 6-OHDA lesion, normalized to mean rates at rest (<0.5 cm s^{-1}) before the lesion. **f**, Ca^{2+} event rates (mean ± s.e.m.) before and after 6-OHDA lesion, normalized to pre-lesion values. $^{***}P < 5 \times 10^{-17}$; signed-rank test comparisons to pre-lesion; $n = 12$ *Drd1a*^{cre}, 13 *Adora2a*^{cre} mice. NS, not significant. **g**, Mean Ca^{2+} event rates (top), SPN spatial coordination (middle) and locomotor speed (bottom) relative to motion onset and offset, before and 14 days after 6-OHDA lesion. Given their similarity (Fig. 2c), pre-lesion traces for both SPN-types were combined. Traces are normalized to mean values from −2 s to −1 s relative to motion onset. Motion-related activity decreased in both SPN types after 6-OHDA lesion ($P = 10^{-61}$ (*Drd1a*^{cre}); $P = 10^{-111}$ (*Adora2a*^{cre})), but was more profoundly reduced in iSPNs. Rank-sum test; $n = 1{,}905$–$3{,}110$ motion bouts for both mouse lines at both time points. Spatial coordination of iSPN ($P = 10^{-9}$) but not dSPN activity ($P = 0.4$) decreased on day 14 to below pre-lesion values. $n = 1{,}326$–$2{,}957$ bouts with ≥2 active cells, for both mouse lines and time points; rank-sum test. **h**, Mean speeds relative to occurrences of Ca^{2+} events, normalized to values 2 s before Ca^{2+} excitation. At 14 days after 6-OHDA lesion, the mean speed rise within ±500 ms of a Ca^{2+} event was 0.8 ± 0.3% (mean ± s.e.m.) for iSPNs, 11 ± 1% for dSPNs. $n = 3{,}498$–$5{,}529$ cells per day, for both mouse lines. **i**, Rises in Ca^{2+} event rates at motion onset, based on data from **g**. For each group, we averaged over all movement bouts and compared the event rates during the intervals −2 to −1 s before and 0–1 s after motion onset. Values normalized to the pre-lesion values. $^{*}P = 0.03$; $^{**}P = 9 \times 10^{-7}$; rank-sum test compared to pre-lesion values; $n = 12$ *Drd1a*^{cre} and 13 *Adora2a*^{cre} mice. **j**, Coactivity values in proximal SPN pairs (20–100 μm apart) during locomotion, after subtraction of coactivity values in shuffled datasets. Values normalized to pre-lesion values. $^{**}P < 2 \times 10^{-4}$; $^{***}P < 4 \times 10^{-12}$; rank-sum test comparing pre-lesion to 1 day and >14 days; $n = 96$–416 coactivity values for both mouse lines and all three time points. **c**, **e**–**j**, Data are from the same 12 *Drd1a*^{cre} and 13 *Adora2a*^{cre} mice. **e**, **g**, **h**, Colour shading indicates the s.e.m. Pre-lesion data were acquired on day −5. Data for >14 days were from 1-h recordings on day 14 plus 30-min recordings on days 15, 17 and 20 performed after saline vehicle administration but before drug treatment.

6-OHDA infusion, we successively administered 1 or 6 mg kg^{-1} of a D2R agonist (quinpirole), a D1R agonist (SKF81297), or L-DOPA. As in past work using the 6-OHDA model of Parkinson's disease, these agents induced locomotor turning contralateral to the lesioned SNc[25], to equipotent levels within the tested dosage ranges (Fig. 4a and Extended Data Fig. 9a–e).

All three drugs also increased dSPN activity, unexpectedly so for quinpirole, given the lack of D2Rs on dSPNs[28,29] (Fig. 4b–d). Quinpirole and the 6 mg kg^{-1} L-DOPA dose, but not SKF81297, further ameliorated the activity imbalance by reducing the elevated iSPN activity in resting mice. The two agonists had divergent effects from those in normal mice, in which they both unexpectedly depressed SPN activity (Extended Data Fig. 7d–f and Supplementary Note). Given the unilateral lesion and systemic drug administration, drug effects in the contralateral striatum or even outside striatum could contribute to these observations, although the brain's greatest densities of dopaminergic inputs and dopamine receptors are in the striatum[30].

Beyond rebalancing SPN activity, L-DOPA mostly restored the coupling to locomotion and spatial clustering of iSPN activity (Fig. 4b, e–i). D2R agonism failed to restore the iSPN activity rise at motion onset, whereas D1R agonism rescued neither the elevated rates nor spatial clustering of iSPN activity (Fig. 4c, f–i). Thus, although L-DOPA and both agonists reversed the activity imbalance, only L-DOPA reversed the deficits in both motion-related iSPN activity and its spatial coordination. These differences in treating neural abnormalities arose even at doses equally affecting turning behaviour, the standard metric of therapeutic efficacy in this model of Parkinson's disease[25].

Late in the treatment of Parkinson's disease, patients commonly develop LID, which we modelled by giving parkinsonian mice a dyskinesiogenic (10 mg kg^{-1}) L-DOPA dose[31] that induced abnormal facial, limb and trunk movements (Extended Data Fig. 9f–j). During dyskinesia, dSPNs were hyperactive and iSPNs hypoactive. Furthermore, dSPNs lost their spatial coordination and movement-evoked responses, whereas iSPNs became more responsive to motion onset (Fig. 5). Thus, the signature dSPN and iSPN dynamics in parkinsonian and dyskinetic states were diametric opposites of each other (Supplementary Video 4). This anti-symmetry provides an attractive explanation for the opposite motor symptoms of Parkinson's disease and LID.

Implications for basal ganglia function

Our data show dSPNs and iSPNs normally have indistinguishable kinetics at motion onset and offset[12] (Fig. 2), challenging the idea that the two cell types respectively mark movement start and end[32]. Activation patterns in the two SPN types were comparably clustered, consistent with previous reports[13,14]. Supporting one recent study[13], the clusters of active SPNs that we identified were neither absolutely stereotyped nor spatially compact.

The ensemble activity that we observed in both SPN types precisely encoded locomotor speed (Extended Data Fig. 4l, m) and as in recent work[13] arose in clusters that preferentially encoded specific movement types (Extended Data Fig. 3, 10). Simultaneous Ca^{2+} imaging in both SPN types revealed interspersed patterns of dSPN and iSPN activity (Extended Data Fig. 4j, k). Together, these findings pose a challenge to 'suppression–selection' models[15], in which the main iSPN function is to broadly suppress competing actions other than the one selected. Such models predict that iSPN dynamics should be less selective, more diffuse and perhaps less velocity-sensitive than those of dSPNs. Instead,

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

it appears that iSPNs and dSPNs jointly engage in movement selection via synchronized activation in interspersed cell clusters of comparable size coupled indistinguishably to locomotion. The data here come from the dorsomedial striatum, and future work should explore whether the findings generalize to other striatal regions.

The spatial extent of dSPN and iSPN coactivation resembles that of corticostriatal neurons with focal axonal arborizations (around 100–300 μm wide) within the rodent striatal matrix[33]. Striatal areas receiving these projections are termed matrisomes and have somatotopically organized cortical inputs and projections to basal ganglia output nuclei[34]. Thus, the spatiotemporal organization of cortical input and the convergence of signals from the direct and indirect pathways are probably both vital for sculpting the motor output signals of the basal ganglia[15,32,35] (Extended Data Fig. 11).

Signatures of a parkinsonian state

The notion that an excess or deficiency of dopamine signalling unbalances striatal dynamics has long provided the conceptual basis for treating basal ganglia disorders[1,2,5–10], but direct evidence of imbalanced striatal activity has been lacking. We uncovered opposite imbalances in dorsomedial striatum after dopamine depletion and LID, and additional facets of SPN pathophysiology that probably contribute to the symptoms of Parkinson's disease, their reversal by dopamine restoration, and the hyperkinesia of LID.

We found four distinct signatures of a parkinsonian state: reduced spontaneous dSPN activity; increased spontaneous iSPN activity; reduced coupling of iSPN activity levels to motion onset and offset; and decreased spatial clustering of iSPN activity (Extended Data Fig. 11). Collectively, these signatures expose the incompleteness of the rate model. Notwithstanding, these findings support the model's predictions for how spontaneous rates of SPN activity change in Parkinson's disease. Abnormal rates of spontaneous SPN activity may indeed impede movement by increasing activity in the inhibitory output projections of the basal ganglia. The loss of spatially clustered iSPN activity and its coupling to movement lie beyond the rate model and probably also contribute to parkinsonian symptoms. These unexpected deficits might indicate disruptions to matrisome[33,34] function that preclude the spatiotemporal dynamics required in basal ganglia output nuclei for well-choreographed actions. The added deficits did not arise after acute dopamine depletion or D1R or D2R antagonism and are consistent with electrophysiological recordings in a genetic mouse model of progressive dopamine cell loss[36].

These progressive pathophysiologic changes probably reflect neural adaptations that occur during chronic dopamine depletion, many of which preferentially affect iSPNs, such as reduced dendritic spine density and increased input from inhibitory interneurons (Supplementary Note). Notably, the long-term adaptations in SPN dynamics were precisely those best treated by L-DOPA and D2R agonism. The mechanisms by which these treatments immediately counteracted the long-term facets of SPN pathophysiology are unclear but important to identify, because the capabilities of the different compounds for reversing the neural abnormalities matched their clinical utilities[18] (in order of efficacy: L-DOPA, D2R agonism, D1R agonism) and might underlie the varying utility of dopamine agonist monotherapy versus that of L-DOPA in different stages of Parkinson's disease[37].

Anti-symmetric deficits in LID

Notably, during LID the dorsomedial striatum exhibits diametric abnormalities to those in the parkinsonian state. Dyskinesia is characterized by decreased spontaneous iSPN activity, increased spontaneous dSPN activity and a loss of spatially clustered and movement-coupled dSPN activity. These features are the polar opposite of those of the parkinsonian state and may explain the opposite motor symptoms of the two conditions. They also fit with reported cellular adaptations in Parkinson's disease versus LID, including reductions of dendritic spines on iSPNs and dSPNs, respectively[19,27]. A caveat is that we examined a model of acute LID, which differs from models with more prolonged

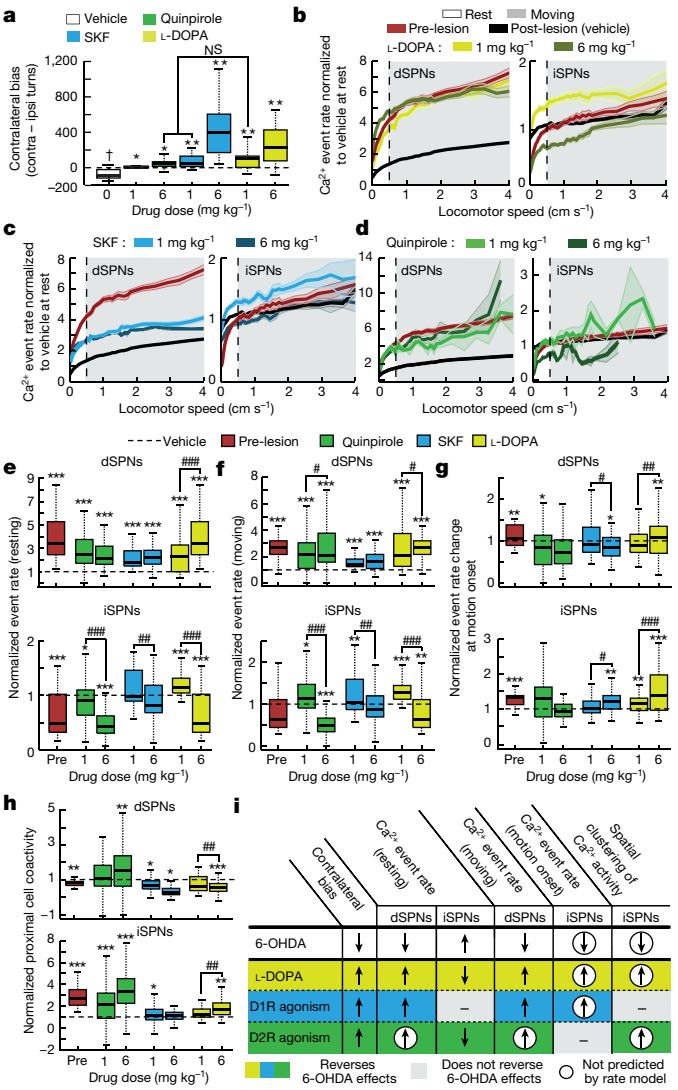


Fig. 4 | L-DOPA better reversed the deficits in SPN activity than dopamine receptor agonism. a, Contralateral turning bias during 40-min periods of locomotion, >14 days after 6-OHDA lesion, after administration of saline vehicle, or 1 mg kg^{-1} or 6 mg kg^{-1} D2R agonist (quinpirole), D1R agonist (SKF81297) or L-DOPA. $n = 13$ mice; one treatment per day, in the order stated. **b**–**d**, Mean SPN Ca^{2+} event rates versus locomotor speed, under different treatment conditions (days 15–21 after 6-OHDA lesion). Values were normalized to mean event rates in mice at rest (<0.5 cm s^{-1}) after saline injection. Colour shading indicates s.e.m. **b**–**h**, Data are from the same 7 *Drd1a*^{cre}, 6 *Adora2a*^{cre} mice. **e**, **f**, Ca^{2+} event rates when mice were resting (**e**) or moving (**f**). Values in **e**–**h** were normalized for each mouse to measured values after saline treatment and 6-OHDA lesion (dashed lines). $n = 14$ speed bins <0.5 cm s^{-1}, $n = 24$ speed bins >0.5 cm s^{-1}, $n = 7$ *Drd1a*^{cre}, 6 *Adora2a*^{cre} mice. **g**, Increases in Ca^{2+} event rates at motion onset. For each treatment, we averaged over all motion bouts and computed the mean rate 0–2 s after motion onset, normalized to rates from −2 to −1 s before motion onset (dashed lines). $n = 11$ time bins per mouse per treatment. **h**, Coactivity in proximal SPN pairs (20–100 μm apart) during movement, after subtracting coactivity values in shuffled datasets. $n = 8$ bins of cell–cell separation, $n = 7$ *Drd1a*^{cre}, 6 *Adora2a*^{cre} mice. **i**, Summary of effects. Arrows indicate mean changes of ≥10% at the 6 mg kg^{-1} dose of each drug. Encircled arrows indicate effects unpredicted by the rate model of Parkinson's disease. Statistical tests are signed-rank tests with Dunn–Sidak correction for multiple comparisons. †$P = 2.4 \times 10^{-3}$ for comparisons to pre-lesion; *$P < 6 \times 10^{-3}$, **$P < 6 \times 10^{-4}$, ***$P < 6 \times 10^{-7}$ for comparisons to vehicle treatment; #$P < 0.05$, ##$P < 5 \times 10^{-3}$, ###$P < 5 \times 10^{-7}$ for comparisons between drug doses. Pre-lesion data were acquired on day −5. Vehicle data were aggregated for 30-min recordings on days 15–21 after saline administration but before drug treatment.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

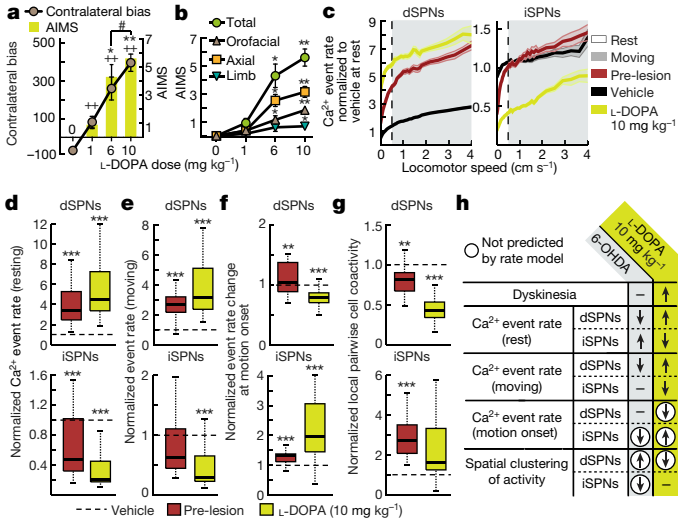


Fig. 5 | In LID, iSPNs are hypoactive and dSPNs have uncorrelated patterns of hyperactivity. a, Turning bias (mean ± s.e.m.) and abnormal involuntary movement score (AIMS) within 40 min after saline vehicle or L-DOPA injection, >14 days after 6-OHDA lesion. $\#\#P < 10^{-3}$ for comparisons of contralateral bias to vehicle treatment; signed-rank test; $*P = 10^{-2}$ and $**P = 10^{-4}$ comparing AIMS values to those after vehicle treatment; rank-sum test; $n = 13$ mice. 10 mg kg^{-1} L-DOPA induced higher AIMS values than 6 mg kg^{-1}. $\#P = 0.013$; signed-rank test. **b**, Peak AIMS (mean ± s.e.m.), separated by movement type, within 40 min after vehicle or L-DOPA treatment in mice >14 days after 6-OHDA lesion. $*P < 0.05$, $**P < 5 \times 10^{-3}$ for comparisons to vehicle treatment; rank-sum test; $n = 13$ mice. **c**, Mean SPN Ca^{2+} event rates versus locomotor speed before 6-OHDA lesion, and >14 days after the lesion following either vehicle or 10 mg kg^{-1} L-DOPA treatment. Values are normalized to mean rates when mice were resting (<0.5 cm s^{-1}) after vehicle treatment. Grey shading indicates >0.5 cm s^{-1} locomotion. Colour shading indicates the s.e.m. **c**–**g**, Data are from the same 7 *Drd1a*^{cre}, 6 *Adora2a*^{cre} mice. **d**, **e**, Ca^{2+} event rates when mice were resting (<0.5 cm s^{-1}; **d**) or moving (>0.5 cm s^{-1}; **e**). All values in **d**–**g** are normalized to those after vehicle treatment in lesioned mice (dashed lines). Statistical tests in **d**–**g** are signed-rank tests with Dunn–Sidak correction for multiple comparisons. $*P < 6 \times 10^{-3}$, $**P < 6 \times 10^{-4}$, $***P < 6 \times 10^{-7}$ for comparisons to vehicle treatment. $n = 14$ speed bins <0.5 cm s^{-1}, 24 speed bins >0.5 cm s^{-1} for all 13 mice. **f**, Rises in Ca^{2+} event rates at motion onset. For each treatment, we averaged over all motion bouts and computed the mean rate 0–2 s after motion onset, relative to the values from −2 to −1 s before onset. $n = 11$ time bins per mouse per treatment. **g**, Coactivity in proximal SPN pairs (20–100 μm apart) during movement, after subtracting coactivity values in shuffled datasets. $n = 8$ bins of cell–cell separation per mouse. **h**, Summary of effects. Arrows indicate mean changes ≥10%. Encircled arrows indicate effects unpredicted by the classical rate model. Pre-lesion data were acquired on day −5. Vehicle data aggregates 30-min recordings on days 15–21 were acquired after vehicle administration but before drug treatment.

courses of L-DOPA treatment known to induce a range of cellular adaptations[27].

As with uncorrelated iSPN hyperactivity in the parkinsonian state, uncorrelated dSPN hyperactivity in LID probably also disrupts movement selection, but in this case causing excessive, unplanned movements. However, it remains unclear whether and how dSPN activity patterns relate to specific dyskinetic movements. Just as pharmacological restoration of spatially coordinated iSPN activity was associated with clinical efficacy in treating the symptoms of Parkinson's disease, reversing or preventing the loss of spatially coordinated dSPN activity in LID might benefit patients with late-stage Parkinson's disease. Overall, the distinctive neural signatures of Parkinson's disease and LID established here may provide biomarkers facilitating development of substitutes for L-DOPA or treatment regimens extending its period of efficacy.

Beyond Parkinson's disease and dyskinesia, our results provide the impetus and means for testing whether other basal ganglia disorders unbalance striatal activity and disrupt its coding dynamics[5,7–10]. As with Parkinson's disease and LID, it is probable that ideas beyond the rate model framework will be required to explain the spectrum of basal ganglia abnormalities and patient symptoms. Current treatment approaches often aim to correct activity imbalances, but our results indicate that next-generation therapies should also target the concomitant biological adaptations that disrupt striatal neural ensemble coding.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0090-6.

Received: 30 May 2016; Accepted: 27 March 2018;
Published online 2 May 2018.

1. Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. *Trends Neurosci.* **12**, 366–375 (1989).
2. DeLong, M. R. Primate models of movement disorders of basal ganglia origin. *Trends Neurosci.* **13**, 281–285 (1990).
3. Shen, W., Flajolet, M., Greengard, P. & Surmeier, D. J. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science* **321**, 848–851 (2008).
4. Mallet, N., Ballion, B., Le Moine, C. & Gonon, F. Cortical inputs and GABA interneurons imbalance projection neurons in the striatum of parkinsonian rats. *J. Neurosci.* **26**, 3875–3884 (2006).
5. Castellan Baldan, L. C. et al. Histidine decarboxylase deficiency causes Tourette syndrome: parallel findings in humans and mice. *Neuron* **81**, 77–90 (2014).
6. Jenner, P. Molecular mechanisms of l-DOPA-induced dyskinesia. *Nat. Rev. Neurosci.* **9**, 665–677 (2008).
7. Lobo, M. K. et al. Cell type-specific loss of BDNF signaling mimics optogenetic control of cocaine reward. *Science* **330**, 385–390 (2010).
8. Schwartz, N. et al. Decreased motivation during chronic pain requires long-term depression in the nucleus accumbens. *Science* **345**, 535–542 (2014).
9. Francis, T. C. et al. Nucleus accumbens medium spiny neuron subtypes mediate depression-related outcomes to social defeat stress. *Biol. Psychiatry* **77**, 212–222 (2015).
10. Moore, H., West, A. R. & Grace, A. A. The regulation of forebrain dopamine transmission: relevance to the pathophysiology and psychopathology of schizophrenia. *Biol. Psychiatry* **46**, 40–55 (1999).
11. Kravitz, A. V. et al. Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature* **466**, 622–626 (2010).
12. Cui, G. et al. Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature* **494**, 238–242 (2013).
13. Klaus, A. et al. The spatiotemporal organization of the striatum encodes action space. *Neuron* **95**, 1171–1180 (2017).
14. Barbera, G. et al. Spatially compact neural clusters in the dorsal striatum encode locomotion relevant information. *Neuron* **92**, 202–213 (2016).
15. Mink, J. W. The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* **50**, 381–425 (1996).
16. Galvan, A. & Wichmann, T. Pathophysiology of Parkinsonism. *Clin. Neurophysiol.* **119**, 1459–1474 (2008).
17. Bronfeld, M. & Bar-Gad, I. Loss of specificity in basal ganglia related movement disorders. *Front. Syst. Neurosci.* **5**, 38 (2011).
18. Jankovic, J. & Aguilar, L. G. Current approaches to the treatment of Parkinson's disease. *Neuropsychiatr. Dis. Treat.* **4**, 743–757 (2008).
19. Cenci, M. A. & Konradi, C. Maladaptive striatal plasticity in l-DOPA-induced dyskinesia. *Prog. Brain Res.* **183**, 209–233 (2010).
20. Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
21. Gong, S. et al. Targeting Cre recombinase to specific neuron populations with bacterial artificial chromosome constructs. *J. Neurosci.* **27**, 9817–9823 (2007).
22. Ghosh, K. K. et al. Miniaturized integration of a fluorescence microscope. *Nat. Methods* **8**, 871–878 (2011).
23. Ziv, Y. et al. Long-term dynamics of CA1 hippocampal place codes. *Nat. Neurosci.* **16**, 264–266 (2013).
24. Barretto, R. P. et al. Time-lapse imaging of disease progression in deep brain areas using fluorescence microendoscopy. *Nat. Med.* **17**, 223–228 (2011).
25. Schwarting, R. K. & Huston, J. P. The unilateral 6-hydroxydopamine lesion model in behavioral brain research. Analysis of functional deficits, recovery and treatments. *Prog. Neurobiol.* **50**, 275–331 (1996).
26. Montastruc, J. L., Llau, M. E., Rascol, O. & Senard, J. M. Drug-induced Parkinsonism: a review. *Fundam. Clin. Pharmacol.* **8**, 293–306 (1994).
27. Fieblinger, T. et al. Cell type-specific plasticity of striatal projection neurons in Parkinsonism and l-DOPA-induced dyskinesia. *Nat. Commun.* **5**, 5316 (2014).
28. Gerfen, C. R. et al. D_1 and D_2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science* **250**, 1429–1432 (1990).
29. Gong, S. et al. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917–925 (2003).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

30. Bentivoglio, M. & Morelli, M. in *Handbook of Chemical Neuroanatomy* Vol. 21 (eds Dunnett, S. B. et al.) 1–107 (Elsevier, Amsterdam, 2005).
31. Cenci, M. A. & Lundblad, M. Ratings of l-DOPA-induced dyskinesia in the unilateral 6-OHDA lesion model of Parkinson's disease in rats and mice. *Curr. Protoc. Neurosci.* **41**, 9.25.1–9.25.23 (2007).
32. Hikosaka, O., Takikawa, Y. & Kawagoe, R. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol. Rev.* **80**, 953–978 (2000).
33. Kincaid, A. E. & Wilson, C. J. Corticostriatal innervation of the patch and matrix in the rat neostriatum. *J. Comp. Neurol.* **374**, 578–592 (1996).
34. Graybiel, A. M., Aosaki, T., Flaherty, A. W. & Kimura, M. The basal ganglia and adaptive motor control. *Science* **265**, 1826–1831 (1994).
35. Alexander, G. E. & Crutcher, M. D. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends Neurosci.* **13**, 266–271 (1990).
36. Panigrahi, B. et al. Dopamine is required for the neural representation and control of movement vigor. *Cell* **162**, 1418–1430 (2015).
37. Perez-Lloret, S. & Rascol, O. Dopamine receptor agonists for the treatment of early or advanced Parkinson's disease. *CNS Drugs* **24**, 941–968 (2010).
38. Paxinos, G. & Franklin, K. B. *The Mouse Brain in Stereotaxic Coordinates* (Academic, San Diego, 2001).

Acknowledgements We appreciate support from HHMI, the Stanford CNC Program, Stanford Photonics Research Center, Pfizer and a GG Technologies gift fund; fellowships from Stanford (J.D.M., T.H.K.), the Helen Hay Whitney Foundation (J.D.M.), the US National Institutes of Health (J.G.P., J.B.D.), HHMI (B.A.), the US National Science Foundation (B.A.), the Bill & Melinda Gates Foundation (B.A.), and the Swiss National Science Foundation (B.F.G.). We thank L. Burns, L. Kitch, E. Hamel, J. Lecoq, M. Larkin, T. Fieblinger, S. Ganguli, A. Girasole, A. Graybiel, A. Kreitzer, R. Malenka and A. Nelson for technical assistance and discussion, Inscopix Inc. for technical support and upgrades, and B. Rossi for scientific illustration.

Reviewer information *Nature* thanks D. Surmeier, G. Stuber and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.G.P., J.D.M., M.D.E. and M.J.S. designed experiments and analyses. J.G.P., J.D.M., B.F.G. and T.H.K. established in vivo imaging protocols and performed experiments. J.D.M., J.G.P. and B.A. analysed data. J.Z.L. and Y.Z. constructed the viral vectors. Y.-W.W. and J.B.D. performed and analysed in vitro electrophysiological recordings. J.G.P., J.D.M. and M.J.S. wrote the paper. M.D.E. and all other authors edited the paper. M.D.E. and M.J.S. supervised the research.

Competing interests M.J.S. is a scientific cofounder of Inscopix, Inc., which produces the miniature fluorescence microscope used in this study. M.D.E. and J.G.P. were Pfizer employees during the initial part of the project.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0090-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0090-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.D.E. and M.J.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Mice. We obtained the GENSAT BAC-transgenic Cre-driver mouse lines, *Drd1a*^{cre} (FK150) and *Adora2a*^{cre} (KG139), from the Mutant Mouse Research & Resource Centers (https://www.mmrrc.org). We crossed these mice with wild-type C57BL/6 mice. We inbred the resulting heterozygous offspring to generate homozygous BAC-transgenic animals, which we maintained through further homozygous crosses. To generate mice for our two-photon imaging experiments, we crossed homozygous *Drd1a*^{cre} with homozygous reporter mice (*Ai14* (also known as *Gt(ROSA)26Sor*$^{tm14(CAG-tdTomato)Hze}$ mice); Allen Brain Institute) expressing the red fluorophore tdTomato in a Cre-dependent manner. We used both male and female mice, aged 12–24 weeks at the start of behavioural experiments. The Stanford Administrative Panel on Laboratory Animal Care (APLAC) approved all experiments.

Virus injections. For miniature microscope and brain slice experiments, we used AAV2/9-CAG-FLEX-GCaMP6m-WPRE.SV40 (1.37×10^9 genome copies (GC) μl^{-1}; Penn Vector Core). For two-photon imaging experiments, we used the *Camk2a* promoter to drive GCaMP expression (AAV2/9-CaMKII-GCaMP6m; 1.8×10^9 GC μl^{-1}) in both SPN types. To inject the virus, we anaesthetized mice with isoflurane (2% in O_2) and stereotaxically injected virus into the dorsomedial striatum (anteroposterior (AP): 1.0 mm from bregma; mediolateral (ML): 1.5 mm from midline; dorsoventral (DV): −2.7 mm (down to −3.2 mm then back up to −2.7 mm from dura)). The injection needle had a 33-gauge bevelled tip. We injected 500 nl of virus at a rate of 250 nl min^{-1}. After the injection was done, we waited 10 min and then withdrew the syringe. We recorded the bregma–dura distance for subsequent positioning of the optical guide tube, sutured the scalp, and allowed the mice to recover for a week before implanting the guide tube.

In vitro electrophysiological and Ca^{2+} imaging studies in striatal tissue slices. Using established techniques[39], we performed dual Ca^{2+} imaging and electrophysiological studies of SPN dynamics in live brain slices obtained from *Drd1a*^{cre} and *Adora2a*^{cre} mice >3 weeks after virus injection. The numerical aperture of the imaging system was matched to that of the miniature microscope used in freely behaving mice. We imaged striatal Ca^{2+} transients at a frame acquisition rate of 20 Hz, using a scientific-grade CMOS camera (ORCA-Flash 4.0 LT, Hamamatsu Photonics) and MicroManager software (NIH)[40]. Concurrently, we performed whole-cell current-clamp recordings of SPN activity using a patch pipette (3–5 MΩ) and a Multiclamp 700B (Molecular Devices) patch-clamp amplifier. Additional details can be found in the Supplementary Methods.

Surgeries. To prepare mice undergoing 6-OHDA lesions and imaging, we anaesthetized mice with isoflurane (2% in O_2) and used a 0.5-mm-diameter drill bit to create a craniotomy dorsal to the SNc (AP = −3.5 mm; ML = 1.25 mm) for insertion of a 26-gauge guide cannula (2.5-mm-long projection, Plastics One, Inc.). We drilled three additional small holes (0.5-mm diameter) at spatially distributed locations for insertion of stabilizing skull screws.

We used a 1.4-mm-diameter drill bit to create another craniotomy (AP = 1.0 mm; ML = 1.5 mm) for implantation of the optical guide tube. We fabricated this guide tube by using ultraviolet liquid adhesive (Norland #81) to fix a 2.5-mm-diameter disc of #0 glass to the tip of a 3.8-mm-long, extra-thin 18-gauge stainless steel tube (McMaster-Carr). We ground off any excess glass using a polishing wheel (Ultratec). Using a 27-gauge blunt-end needle, we aspirated the cortex down to DV = −2.1 mm from the dura and implanted the exterior glass face of the optical guide tube at DV = −2.35 mm. We lowered the SNc guide cannula to DV = −1.0 mm from the dura. After stereotaxic placement of these components, we applied Metabond (Parkell) to the skull and fixed the full assembly using dental acrylic (Coltene). After the acrylic hardened, we used a second batch of acrylic to attach an aluminium head bar to the cranium, for the purpose of head-fixing the mice during subsequent experimental procedures. Mice recovered for 3–4 weeks before two-photon imaging experiments or mounting of the miniature microscope.

Mounting of the miniature microscope. At 3–4 weeks after guide tube implantation, we assessed the extent of Ca^{2+} indicator expression in the dorsomedial striatum. To do this, we inserted a gradient refractive index (GRIN) lens (1 mm diameter; 4.12 mm length; 0.46 numerical aperture; 0.45 pitch; GRINTECH GmbH or Inscopix Inc.) into the optical guide tube and used a commercial two-photon fluorescence microscope (Ultima, Bruker) to image the striatum of each mouse while it was anaesthetized with isoflurane (2% in O_2). In mice with uniform indicator expression, we secured the GRIN lens in the guide tube with ultraviolet (UV)-light curable epoxy (Loctite 4305).

After affixing the GRIN lens, we lowered a miniature microscope (nVistaHD, Inscopix Inc.) towards the GRIN lens until the fluorescent tissue was in focus. To secure the miniature microscope to the cranium, we created a base on the cranium around the GRIN lens using blue-light curable resin (Flow-It ALC; Pentron). We attached the base plate of the miniature microscope to the resin base using UV-light curable epoxy (Loctite 4305). After affixing its base plate, we released the microscope and attached a base plate cover (Inscopix Inc.). We coated the resin with black nail polish (Black Onyx, OPI) to make it opaque.

6-OHDA infusions. To block serotonin and noradrenaline transporters and thereby prevent lesioning of non-dopamine monoamine neurons, we injected mice with desipramine (25 mg kg^{-1}; i.p.) 30 min before 6-OHDA infusion[41]. We anaesthetized the mice with isoflurane (2% in O_2) and stereotaxically inserted into the craniotomy above the SNc (see 'Surgeries') an injection cannula (33 gauge, 17-mm-long) that was connected to a syringe pump loaded with 6-OHDA (4 $\mu g\,\mu l^{-1}$ in saline). We sequentially infused 6-OHDA (650 nl at 100 nl min^{-1}) at three sites: DV = −4.2 mm, −4.0 mm and −3.8 mm below the dura. We withdrew the injection cannula 10 min after the final infusion. After infusion, we monitored the body mass of each mouse and subcutaneously administered 1 ml lactated Ringer's solution to prevent dehydration. In the first 3 days after 6-OHDA infusion, we provided the mice with manually crumbled mouse chow, to aid recovery.

Pharmacology. We dissolved the D1R agonist (SKF81297; 1, 3 or 6 mg kg^{-1}), D1R antagonist (SCH23390; 0.2 mg kg^{-1}), D2R agonist (quinpirole; 1, 3 or 6 mg kg^{-1}), D2R antagonist (raclopride; 1 mg kg^{-1}), or L-DOPA (1, 6 or 10 mg kg^{-1}) mixed with benserazide (15 mg kg^{-1}) in 0.9% saline solution (all compounds from Sigma-Aldrich). We delivered all drugs systemically to the mice via intraperitoneal injection (10 ml kg^{-1} injection volume). We selected doses based on published receptor occupancy data in systemically injected rodents[42–44] and our own evaluations of the dose-dependent effects of these drugs on behaviour in hemiparkinsonian mice (Extended Data Fig. 9). Specifically, we aimed to select doses with overlapping ranges of behavioural potency to test during Ca^{2+} imaging (Fig. 4). Note that SCH23390 does bind to 5HT2 receptors, but its receptor occupancy level on 5HT2 receptors is 40–50 times less than that of D1Rs at the dose we used[45,46]. For all Ca^{2+} imaging sessions, we waited 15 min after drug injections before recording neural activity for 30 min. See Extended Data Fig. 9 for the time-dependence of drug efficacy.

Experimental timeline. Before the lesion, we administered mice the D1R antagonist SCH23390 (day −4, 0.2 mg kg^{-1}), the D2R antagonist raclopride (day −3, 1 mg kg^{-1}), the D1R agonist SKF81297 (day −2, 1 mg kg^{-1}), and the D2R agonist quinpirole (day −1, 1 mg kg^{-1}). After the lesion, we administered quinpirole (day 15, 1 mg kg^{-1}; day 16, 6 mg kg^{-1}), SKF81297 (day 17, 1 mg kg^{-1}; day 18, 6 mg kg^{-1}), and the dopamine precursor L-DOPA (day 19, 1 mg kg^{-1}; day 20, 6 mg kg^{-1}; day 21, 10 mg kg^{-1}). On days on which mice received drug treatments, we performed 30 min of Ca^{2+} imaging before administering the drug, and another 30 min of Ca^{2+} imaging afterward.

Ca^{2+} imaging in freely moving mice. Brain imaging in freely moving mice occurred in a circular arena (31 cm in diameter). To habituate mice to this arena, we allowed them to explore it for 1 h on each of three sequential days before any Ca^{2+} imaging. Prior to each imaging session (Fig. 1a), we head-fixed each mouse by its implanted head bar and allowed the mouse to walk or run in place on a running wheel. We then attached the miniature microscope and adjusted the focal setting to optimize the field-of-view, in the first imaging session, or in subsequent sessions to match the field-of-view to that of all prior sessions (Extended Data Fig. 1j).

After securing the microscope to the head of the mouse, we detached the mouse from its head restraint and allowed it to freely explore the circular arena. After allowing ≥10 min for the mouse to habituate to the arena, fluorescence Ca^{2+} imaging commenced using 50–200 μW of illumination power at the specimen plane and a 20-Hz frame acquisition rate. We also acquired 20-Hz movies of the behaviour of each mouse using a TTL-triggered video camera (SVWGAM; EPIX Inc.), a varifocal lens (T3Z2910CS; Computar), an image frame grabber card (PIXCI-SI; EPIX Inc.) and image-acquisition software (XCAP; EPIX Inc.). We concurrently recorded behavioural and Ca^{2+} imaging data in sessions generally lasting 60 min in total. On days on which the mouse received drug treatments (Fig. 1a), we gave the mouse an injection of saline vehicle, waited 5 min, and then performed 30 min of Ca^{2+} imaging. We then administered the drug, waited 15 min for the drug to take effect, and then acquired another 30 min of Ca^{2+} imaging data.

Two-photon Ca^{2+} imaging in behaving mice. We acquired volumetric Ca^{2+} imaging data from four different axial planes in the brains of head-fixed mice free to behave on a running wheel. The mice expressed the red fluorophore tdTomato in dSPNs and GCaMP6m in both SPN types (see 'Mice'). After implantation of an optical guide tube and GRIN microendoscope in the dorsomedial striatum (see 'Surgeries'), we head-fixed the mice beneath the microscope objective lens carrier of a custom-built two-photon microscope[47]. The microscope was equipped with a resonant galvanometer laser-scanning mirror (CRS 8 K, Cambridge Technologies), allowing a 30-Hz frame acquisition rate (512 × 512 pixels) under the control of ScanImage 5.2 software[48] (Vidrio Technologies). Additional details can be found in the Supplementary Methods.

Rotarod assay. Before day −5 (Fig. 1a), we trained mice for 3–4 days on an accelerating rotarod (EzRod; Omnitech) to ensure that they could remain on the rotarod. Each mouse performed three trials per day, with the rod accelerating from 6 to 60 r.p.m. in 3 min. During subsequent rotarod sessions (days −5, 1 and 14), we imaged SPN Ca^{2+} activity as mice walked on the rotarod at a constant speed of 5 r.p.m. for 10 min. After dopamine depletion, we excluded from

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

analyses of SPN Ca^{2+} activity any mouse that could not stay on the rotarod for the full 10 min of Ca^{2+} imaging.

Turning bias assay. To eliminate any potential effects on mouse turning behaviour from the miniature microscope or its cable, we studied the behaviour of a separate cohort of mice that underwent unilateral dopamine depletion. This allowed us to assay the dose-dependent behavioural effects of the same drugs that we studied with in vivo Ca^{2+} imaging. We manually scored the number of rotations that mice made ipsilateral and contralateral to the right cerebral hemisphere during 1 h recordings of open field locomotion. We computed the rotational bias of each mouse as the difference between the total numbers of rotations in each direction. We then lesioned the right SNc with 6-OHDA and repeated this scoring procedure 1 day and 14 days after the lesion.

We then evaluated the effects of intraperitoneal injections of saline vehicle, quinpirole (1, 3 and 6 mg kg^{-1}), SKF81297 (1, 3 and 6 mg kg^{-1}) and L-DOPA (1, 6 and 10 mg kg^{-1}) during 3 h of locomotion in the open field, on sequential days. We manually scored the rotational bias of each mouse off-line in 20-min time bins (Extended Data Fig. 9a–d).

Dyskinesia assay. In the same mice and behavioural sessions used to quantify the turning bias, we measured the abnormal involuntary movement score (AIMS) for each mouse following established procedures[31]. This score is a well-known metric used to quantify dyskinesia levels in parkinsonian mice in response to a range of drug treatments and experimental manipulations[31]. In brief, for a period of 3 h after systemic drug injection, every 20 min we assigned each mouse an AIMS value by observing the behaviour of the mouse for a sample period of 60 s. We monitored the mouse for the presence of abnormal (i) axial, (ii) limb and (iii) orofacial movements. To provide multiple viewing perspectives of the mouse during AIMS scoring, we placed a pair of rectangular mirrors (30.5 cm × 45.7 cm) outside the open field arena, at 90° to each other, behind the mouse and opposite to the experimenter. During each 60-s sample period, we assigned the mouse a score of 0–4 in each category (i–iii). A score of '0' signified no observed abnormalities. A score of '1' or '2' signified abnormal behaviour that respectively lasted <30 s or >30 s. A '3' signified abnormal behaviour that lasted 60 s but that could be interrupted by an external stimulus (for example, a loud clap). A '4' indicated abnormal behaviour that lasted 60 s and could not be interrupted. For each 60-s sample period, we summed the individual category AIMS values to attain the total AIMS value.

Histology. After all behavioural and Ca^{2+} imaging experiments, we euthanized and intracardially perfused the mice with phosphate-buffered saline (PBS) and then a 4% solution of paraformaldehyde in PBS. We sliced the fixed brain tissue using a vibratome (Leica VT1000s) to obtain 80-μm-thick coronal sections. We immunostained the tissue sections with antibodies against tyrosine hydroxylase (1:500, Aves TYH) and GFP (1:1,000, Molecular Probes A-11122) and then applied fluorophore-conjugated secondary antibodies (Jackson Immunoresearch 703-586-155 and 711-546-152, respectively). We visualized immunofluorescence in our sections using a slide-scanning microscope (Olympus VS-120) and confirmed that all mice included for post-lesion analyses had a complete unilateral loss of dopamine neurons in the SNc.

Video analysis of mouse behaviour. We used custom software written in MATLAB to analyse the videos (20-Hz frame acquisition rate) of mouse behaviour acquired during Ca^{2+} imaging with the miniature microscope. First, we ranked all the pixels in each video frame according to their intensity values. Under the lighting conditions used to acquire the videos, the mouse appeared much darker than the surrounding open field arena (Supplementary Videos 1, 2). Therefore, we next determined the centroid location of the mouse in each frame as the mean location of the 1,500 darkest pixels. By visual inspection of each movie, we confirmed that this approach reliably tracked the movements of each mouse and yielded an accurate record of its position. To match the temporal resolution of our down-sampled Ca^{2+} recordings, we down-sampled the locomotor speed traces to 5 Hz via linear interpolation. We classified each 200-ms time bin as one in which the mouse was either 'resting' or 'moving', according to the down-sampled speed trace and using a threshold value of 0.5 cm s^{-1} to separate the two states. If the mouse was 'moving' in two time bins separated by <1 s, we classified the intervening time bins as ones in which the mouse was 'moving'.

We applied these analytic procedures uniformly to the behavioural videos across all experimental and drug treatment conditions. Under dyskinetic conditions, mice commonly exhibit ongoing movements of the face, limbs and trunk. In general, these movements did not translate the body centroid sufficiently rapidly to register as locomotor episodes. Nevertheless, an important caveat is that under dyskinetic conditions a greater number of sub-threshold movements may have occurred during the computationally identified 'resting' periods than under normal or parkinsonian conditions.

Classification of movement types. To study and compare movements of different types made by freely moving mice (Extended Data Figs. 3, 8), we used custom software written in MATLAB (Mathworks) to extract from the behavioural videos the intervals from –4 to 4 s surrounding the onset of each movement bout. Using this software, we visually scored each movement bout as an instance of forward locomotion, a left or right turn, grooming, or upward rearing. Additional details are shown in Extended Data Fig. 3.

Quantification of movement in head-restrained mice. To analyse the locomotor behaviour of mice during two-photon Ca^{2+} imaging sessions (see 'Two-photon Ca^{2+} imaging in behaving mice'; Extended Data Fig. 4), we computed the running speed of the mouse using the signals from a rotary encoder on the running wheel and identified movement periods during which the running speed was above a minimum threshold. Additional details are shown in Extended Data Fig. 4.

Basic processing of in vivo fluorescence Ca^{2+} movies. To reduce processing times, we spatially down-sampled the raw Ca^{2+} videos from the miniature microscope using 4 × 4 bi-linear interpolation. To correct images for lateral motions of the brain, we used the image-alignment algorithm, TurboReg[49], to perform a rigid image registration across all frames of the movie. To correct for fluctuations in background fluorescence intensity, we applied a Gaussian low-pass filter to each image, which smoothed all structures <100 μm in size, and then divided each image frame by its low-pass filtered version. Using the resultant normalized movie ($F(t)$), we subtracted the time-averaged, mean fluorescence intensity (F_0) of each pixel to obtain the difference movie, $\Delta F(t) = F(t) - F_0$. We then divided this difference by F_0 to obtain movies of relative fluorescence changes, $\Delta F(t)/F_0$. Finally, we temporally down-sampled (4×) the $\Delta F(t)/F_0$ movies to a 5-Hz frame rate via a linear interpolation, and we used the resulting videos for extraction of individual neurons and their Ca^{2+} activity traces.

We processed the raw two-photon Ca^{2+} videos with the same procedures, but with two exceptions: we did not down-sample the movies in space or time, and we applied a median spatial filter (6 × 6 pixels) to the two-photon $\Delta F(t)/F_0$ movies before proceeding to cell sorting. For display purposes only (Supplementary Video 3), we converted the fluorescence trace, $F(t)$, of each pixel into a trace of z scores, $\Delta F(t)/\sigma$. Here $\Delta F(t) = F(t) - F_0$ denotes the deviation of each pixel from its mean value, F_0, and σ denotes the background noise of each pixel, which we estimated by taking the standard deviation values calculated with a sliding 250-s window.

Identification of individual neurons. We identified individual cell bodies in the brain imaging videos by applying to the pre-processed one- and two-photon $\Delta F(t)/F_0$ movies an established algorithm for cell sorting based on principal component and independent component analyses (PCA–ICA)[50]. As in previous work[50], the extracted spatial filters generally had sizes, morphologies and corresponding Ca^{2+} activity traces that were characteristic of individual neurons, but there were also some filters that were obviously not neurons that we discarded. We verified every cell included in the analyses by visual inspection. We did not register individual cells over multiple days; hence all analyses characterizing the rates, amplitudes and durations of Ca^{2+} events (Extended Data Fig. 1l–n) were conducted separately for each day of experimentation, over the uniform imaging duration (1 h per session).

Cells identified in the two-photon Ca^{2+} videos underwent further analyses, to classify each cell as being either an iSPN or dSPN, and to remove duplicate copies of individual cells found in more than one image plane. We first manually identified the centroids of all visible cells in each plane labelled with tdTomato. We then compared these centroid locations to those of the cells extracted from the two-photon Ca^{2+} videos and matched each tdTomato-expressing cell to the nearest unassigned cell in the Ca^{2+} video (Extended Data Fig. 4b, c).

To remove duplicate copies of individual cells seen on more than one plane of the two-photon Ca^{2+} videos, we computed for all nearby cells found on different planes both the distance between the centroids of the two cells and the Pearson correlation coefficient of their $\Delta F(t)/F_0$ activity traces. We identified duplicate copies of the same cell as cell pairs within 15 μm of each other and with an activity correlation coefficient >0.7. Among the duplicates, we retained the one with the Ca^{2+} activity trace with the highest signal-to-noise ratio.

Detection of Ca^{2+} transients. After identifying individual cells in the one- and two-photon brain imaging datasets, we detected Ca^{2+} events in the fluorescence trace of Ca^{2+} activity of each cell using a threshold-crossing algorithm to detect fluorescence peaks that were ≥ 3 s.d. above the baseline signal value for ≥ 1.4 s. We determined the time of each Ca^{2+} event as the temporal midpoint between the time of the Ca^{2+} event's fluorescence peak and the most recent preceding trough in fluorescence[23].

On 6.4 ± 3% (mean ± s.d.) of all detected Ca^{2+} events in the one-photon datasets, bright fluorescence increases from a Ca^{2+} event in one cell occupied more pixels than a single spatial filter. To prevent contamination of the time traces of other cells we took a conservative approach and allowed only a single cell among a group of neighbours to register a Ca^{2+} event in a single 200-ms time frame. We defined neighbours as cells that had centroids within 20 μm of one another[23]. If multiple Ca^{2+} events arose within 200 ms in neighbouring cells, we retained only the transient with the greatest peak value.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

To characterize the waveforms of the Ca^{2+} transients of the cells (Extended Data Figs. 1l–n, 5f–h), we took the spatial filter of each cell, zeroed all pixels with values <50% of the peak value, and applied the truncated filter to the movie of $\Delta F(t)/F_0$ to generate a time trace of Ca^{2+} activity (Fig. 1d). Using the previously identified Ca^{2+} events, we then considered the statistical distribution of Ca^{2+} transient waveforms among all events and SPNs of a given type. From this distribution, we computed the median and interquartile range values of Ca^{2+} transient amplitude and duration (Extended Data Fig. 1l–n). With the exception of these basic characterizations of the Ca^{2+} transient waveforms, all subsequent analyses made use only of the occurrence times of the Ca^{2+} transients and treated these occurrences as binary events.

Determinations of coactivity levels in cell pairs. SPN ensembles displayed prominent temporal correlations between cells across timescales of ~1–2 s (Supplementary Videos 1–4). To characterize correlations at these timescales, we first applied a 'forward-smoothing' by taking the binarized traces of Ca^{2+} events of individual cells and extending the number of time bins (200 ms per bin after temporal down-sampling; see 'Basic processing of in vivo fluorescence Ca^{2+} movies') in which a cell was marked as active so as to include all frames within 1 s after the Ca^{2+} event. Then, to assess the extent to which pairs of cells had correlated patterns of activity, for each cell pair we computed the Jaccard index characterizing the similarity between the binarized Ca^{2+} activity traces of two cells. We used this index as a metric of pairwise coactivation and determined it as the total number of time bins during which both cells in a pair were active, divided by the total number of time bins in which either cell was active. We chose the Jaccard index because, unlike more common measures of correlation, such as the Pearson correlation coefficient, the Jaccard index does not depend on the number of time bins in which neither cell is active and in this sense is more sensitive to the periods of joint activity rather than those of joint inactivity. We computed the mean value of the Jaccard index as a function of the distance between the cells in a pair (Fig. 2d and Extended Data Figs. 4j, 6e).

To control for changes in pairwise coactivation that might be simply owing to the changes in Ca^{2+} event rates resulting from the experimental manipulations and drug treatments, we also computed the Jaccard index for shuffled datasets in which the Ca^{2+} events of each cell were randomly re-distributed in time. We performed 1,000 independent shuffles of the Ca^{2+} event times for each cell, when mice were resting or in locomotion. To compare pairwise coactivation levels in the real data against those in the shuffled data, we normalized the plots of mean Jaccard index for the real data by the corresponding values for the shuffled data (Fig. 2d, e and Extended Data Figs. 4j, k, 6e). Alternatively, we subtracted the coactivity values determined from the shuffled datasets from those of the real data, and compared the resultant values across different experimental conditions for the proximal cell pairs (cell centroid separations between 20–100 μm lateral separation) (Figs. 3j, 4f, 5g and Extended Data Fig. 7i, j).

We also checked whether SPN activity rates varied across the imaging field-of-view and thus affected estimates of coactivity levels. We computed the Pearson correlation coefficient between the Ca^{2+} event rate of every cell and its distance from the centre of the field-of-view. We found a very weak but statistically significant correlation between these two variables ($r = 0.07 \pm 0.02$; $P < 10^{-10}$; $n = 29{,}097$ cells; $n = 19$ mice, each sampled during five imaging sessions), which explained almost none of the variations in the Ca^{2+} activity rates of the cells across the SPN population ($R^2 = 0.005$). Overall, effects arising from the locations of the cells in the field-of-view were far too weak to account for the substantial distance-dependence of pairwise coactivity levels of the cells (Fig. 2d).

We determined pairwise coactivity levels in the two-photon Ca^{2+} videos using identical procedures as for the one-photon Ca^{2+} videos.

Spatial coordination index. In the Ca^{2+} videos, there were plainly evident, spatially correlated patterns of SPN activation at the seconds time scale. To characterize the spatial structure of these patterns at individual time points in the videos, we computed an index of spatial coordination. To calculate how this index varied over an entire imaging session (Extended Data Fig. 2), we took the binarized traces of the Ca^{2+} events of individual cells and extended the number of time bins (200 ms each after temporal down-sampling; see 'Determinations of coactivity levels in cell pairs') in which a cell was marked as active so as to include all frames within 1 s after each Ca^{2+} event. We performed this temporal expansion, which increased the time window over which cells appeared as jointly active, to highlight the spatially clustered activation of cells at the seconds time scale. For each time bin, we then computed the set of distances between all pairs of cells that were active in the time bin. For each time bin, we also calculated this set of pairwise distances for shuffled versions of the data, in which we randomized the identities of the active cells while maintaining their centroid coordinates as well as the same number of active cells in the time bin. We performed 1,000 independent shuffles for each time bin. Aggregating the results across all 1,000 shuffles, we computed the cumulative distribution function of pairwise distances between active cells. For each time bin, we also computed this cumulative distribution function for the real data, and performed a pair of one-tailed Kolmogorov–Smirnov tests that compared the cumulative distribution functions of pairwise distances in the shuffled and real versions of the data. With one of the two tests, we examined the hypothesis that the active cells in the real data were distributed closer together than expected by chance, and with the other, we examined the hypothesis that the active cells in the real data were distributed further away from each other than expected by chance. We took the logarithm (base ten) of the lesser of the two P values (that is, the P value from the test that more significantly rejected its null hypothesis). We multiplied the resultant number by negative one or one, respectively, according to whether the locations of the active cells in the time bin were spatially correlated or anti-correlated (as determined by which of the two Kolmogorov–Smirnov tests gave the more significant result). We assigned the spatial coordination index (SCI) the value of the resulting product.

To characterize the time-dependent relationship between mouse speed and the SCI, we computed the time-dependent cross-correlation function between the two variables (Extended Data Fig. 2e, f). To analyse the effects of the temporal sampling rate on these calculations, before computing the SCI, we first forward-smoothed (see previous paragraph) the Ca^{2+} event rasters using a variety of different smoothing time constants from 200 to 1,000 ms. After computing the speed-SCI cross-correlation function, we normalized it to the mean of its values between the intervals from −4 to −3 s and from 3 to 4 s.

Analyses of Ca^{2+} event rates as a function of locomotor speed. To examine how the rate of Ca^{2+} events varied with changes in locomotor speed (Figs. 2a, 3e, 4b–d, 5c and Extended Data Figs. 4g, 7d, e), we sorted the individual image frames of the behaviour video of each mouse according to the instantaneous locomotor speed of the mouse. We then binned the image frames into sets of images for which the locomotor speed was the same to within 0.036–0.73 cm s^{-1}, using finer bins at lower speeds and coarser bins at higher speeds. For each image set, we computed its mean rate of Ca^{2+} events across all detected SPNs.

To make comparisons between mice (Fig. 2a), we took the mean Ca^{2+} event rate of each mouse as a function of locomotor speed and normalized this curve to the mean Ca^{2+} event rate of each mouse while it was at rest (speed <0.5 cm s^{-1}). We averaged these normalized traces across the set of all *Drd1a*^{cre} or *Adora2a*^{cre} mice.

To compare Ca^{2+} event rates during the various drug treatments to those following injections of saline vehicle (Figs. 4b–d, 5c and Extended Data Fig. 7d, e), we normalized the Ca^{2+} event rates for each mouse and each drug condition by the mean rate when the mouse was at rest after having been given saline vehicle. For each drug condition, we then averaged these normalized curves across all mice in the cohort.

To compare Ca^{2+} event rates before and after dopamine depletion (Fig. 3e), we first computed the curve of the mean Ca^{2+} event rate of each mouse as a function of locomotor speed. We averaged these curves across mice of a given genotype, and then normalized the resulting trace to the mean rate when the mice were at rest in baseline recordings, before 6-OHDA lesion. For all comparisons, we calculated the s.e.m. for the mean Ca^{2+} event rate for each mouse and speed bin.

Determinations of motion onset- and offset-triggered averages. We identified onsets of locomotion by finding all movement bouts (≥2 s in duration) that occurred ≥2 s after the prior movement bout. We identified offsets of locomotion by finding all movement bouts that preceded the next movement bout by ≥2 s. To calculate motion onset- or offset-triggered averages, for each mouse we first normalized the speed-dependent, mean rates of Ca^{2+} events or mean spatial coordination indices to the corresponding mean values from when each mouse was at rest during the period of −2 to −1 s relative to motion onset. We then took all motion onset and offset occurrences and calculated the time-dependent, onset- and offset-triggered averaged traces of the normalized Ca^{2+} event rate or spatial coordination index. We averaged these traces across all mice in each cohort.

To calculate the average times at which dSPN and iSPN activity reached their half-maximum values relative to the onset of movement, we fit logistic curves to the mean trace of the Ca^{2+} event rate of each mouse for the interval of −2 to 2 s relative to movement onset. We then determined the half-maximum activation time from the fit parameters. We averaged these time values over all mice of the same genotype.

To calculate the average times at which dSPN and iSPN activity commenced relative to the onsets of locomotion on the running wheel (Extended Data Fig. 4h), we calculated the percentage of cells of each type that were active in each time bin in the binarized traces of Ca^{2+} activity (see 'Detection of Ca^{2+} transients'). Across all instances of motion onset, we computed the average number of dSPNs and iSPNs that were active in each time bin over the interval of −3 to 1 s relative to motion onset. We then fit these averaged traces to a logistic curve. From each parameterized fit curve, we determined the onset time as the time at which the mean percentage of active cells was ≥3 s.d. of the percentage of active cells during the baseline periods of −3 to −1 s relative to motion onset.

Analysis of Ca^{2+} event-triggered averages. To examine how the occurrences of the Ca^{2+} events of cells (in the one- or two-photon imaging datasets) correlated in

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

time with changes in the locomotor speed of the mice (Figs. 2b, 3h and Extended Data Fig. 4i), we computed the time trace of the Ca^{2+} event-triggered, mean locomotor speed for each cell. For each mouse, we averaged these event-triggered traces of mean speed across all cells and normalized the resulting trace by its value at either −2 s or −20 s relative to Ca^{2+} event onset, in the one- and two-photon imaging studies, respectively. In experiments with freely moving mice, we averaged these traces across all mice with the same genotype. In dual-colour imaging experiments with head-fixed mice, we averaged the traces across all SPNs of the same type. We computed the first temporal moment of the resulting averaged trace across the interval of −2 to 2 s, to determine the temporal offset between peak locomotor speed and the occurrence of an SPN Ca^{2+} event.

Similarities of the neural ensembles activated by movements of different types. To statistically evaluate the relative similarity of the neural ensembles activated during movement bouts of the same or different types, and during baseline periods, we used the Jaccard similarity index (Extended Data Figs. 3d, e, 8d–g). Additional details can be found in the Supplementary Methods.

Decoding mouse locomotor speed based on SPN Ca^{2+} activity. We used a generalized linear model (GLM) to make time-dependent predictions of the locomotor speeds of mice based on the $\Delta F(t)/F_0$ activity traces of the cells as determined by two-photon Ca^{2+} imaging. Additional details are shown in Extended Data Fig. 4.

Computational simulations of SPN activity traces. To examine the temporal accuracy with which we could use Ca^{2+} events in SPNs to determine the times of the accompanying action potentials, we created and analysed artificial datasets based on the fluorescence responses to action potentials observed in our dual electrical–optical recordings in vitro and on basic facets of the fluorescence recordings in vivo. Additional details can be found in the Supplementary Methods.

Data analyses and statistical tests. We performed data analysis using custom software written in MATLAB and ImageJ. We also used the Prism (Graphpad) and R programming environments for statistical analyses. We used two-tailed, non-parametric statistical tests to avoid assumptions of normal distributions and equal variance across groups. All signed-rank tests were Wilcoxon signed-rank tests. All rank-sum tests were Wilcoxon rank-sum tests. Analyses were performed unblinded. The sizes of our mice samples were chosen to approximately match those of previous work, as there was no pre-specified effect size. In all box-and-whisker plots, horizontal lines denote median values, boxes cover the middle two quartiles, and whiskers span 1.5× the interquartile range. *P* and *n* values for the statistical tests performed in all the figures can be found in the Supplementary Information.

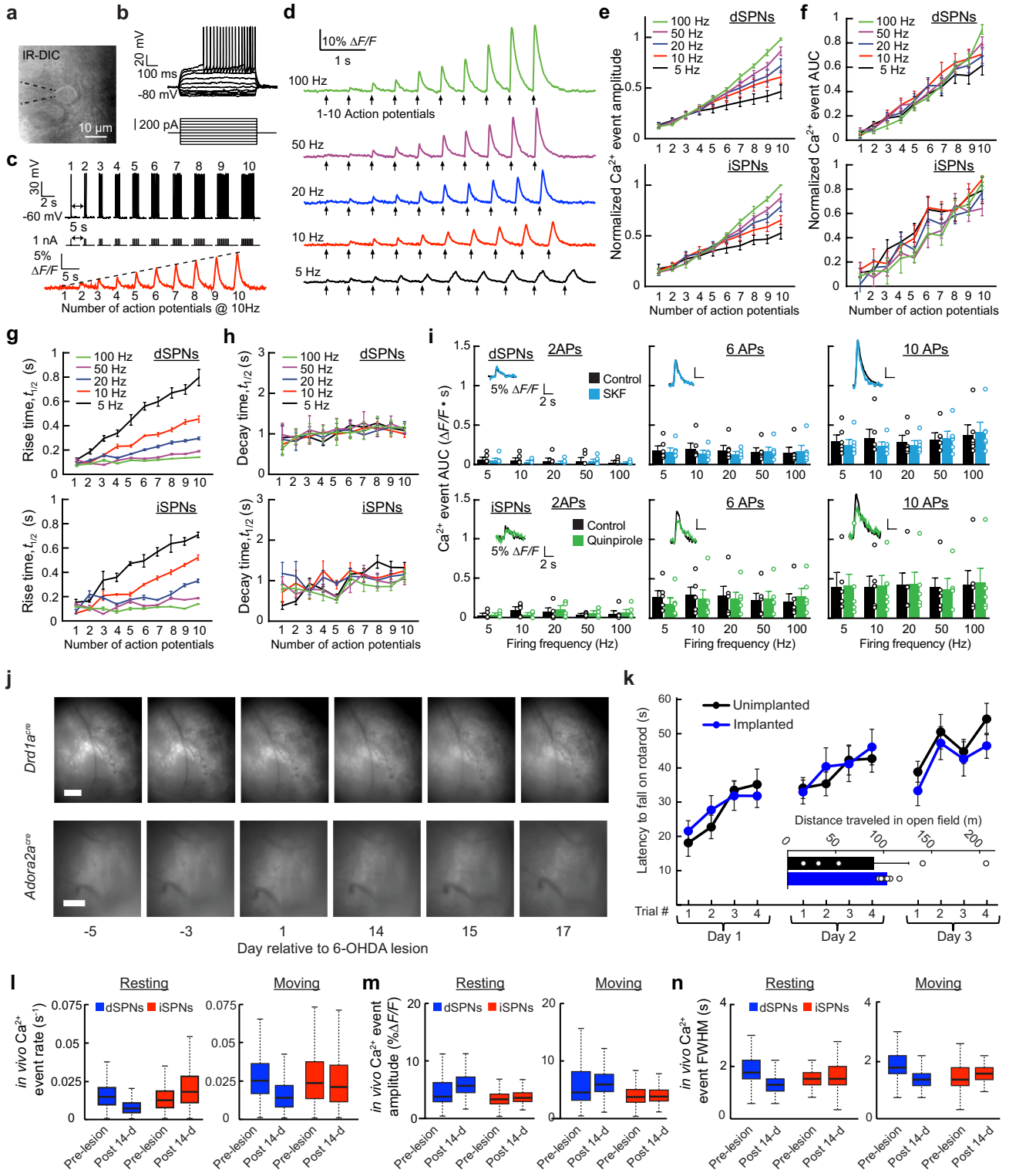
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The algorithm used for image registration is available on its author's website[49]. The algorithm used for cell sorting is available as Supplementary Material published elsewhere[50]. Other software code that support the findings of this study are available from the corresponding authors upon reasonable request.

Data availability. The data that support the findings of this study are available from the corresponding authors upon reasonable request.

39. Wu, Y. W. et al. Input- and cell-type-specific endocannabinoid-dependent LTD in the striatum. *Cell Rep.* **10**, 75–87 (2015).
40. Edelstein, A. D. et al. Advanced methods of microscope control using μManager software. *J. Biol. Methods* **1**, e10 (2014).
41. Breese, G. R. & Traylor, T. D. Depletion of brain noradrenaline and dopamine by 6-hydroxydopamine. *Br. J. Pharmacol.* **42**, 88–99 (1971).
42. Peng, T. et al. D_2 receptor occupancy in conscious rat brain is not significantly distinguished with [^{3}H]-MNPA, [^{3}H]-(+)-PHNO, and [^{3}H]-raclopride. *Synapse* **64**, 624–633 (2010).
43. Neisewander, J. L., Fuchs, R. A., O'Dell, L. E. & Khroyan, T. V. Effects of SCH-23390 on dopamine D_1 receptor occupancy and locomotion produced by intraaccumbens cocaine infusion. *Synapse* **30**, 194–204 (1998).
44. Wadenberg, M. L., Kapur, S., Soliman, A., Jones, C. & Vaccarino, F. Dopamine D_2 receptor occupancy predicts catalepsy and the suppression of conditioned avoidance response behavior in rats. *Psychopharmacology* **150**, 422–429 (2000).
45. Riddall, D. R. A comparison of the selectivities of SCH 23390 with BW737C89 for D_1, D_2 and 5-HT_2 binding sites both in vitro and in vivo. *Eur. J. Pharmacol.* **210**, 279–284 (1992).
46. Suhara, T. et al. D_1 dopamine receptor binding in mood disorders measured by positron emission tomography. *Psychopharmacology* **106**, 14–18 (1992).
47. Lecoq, J. et al. Visualizing mammalian brain area interactions by dual-axis two-photon calcium imaging. *Nat. Neurosci.* **17**, 1825–1829 (2014).
48. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).
49. Thévenaz, P., Ruttimann, U. E. & Unser, M. A pyramid approach to subpixel registration based on intensity. *IEEE Trans. Image Process.* **7**, 27–41 (1998).
50. Mukamel, E. A., Nimmerjahn, A. & Schnitzer, M. J. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* **63**, 747–760 (2009).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

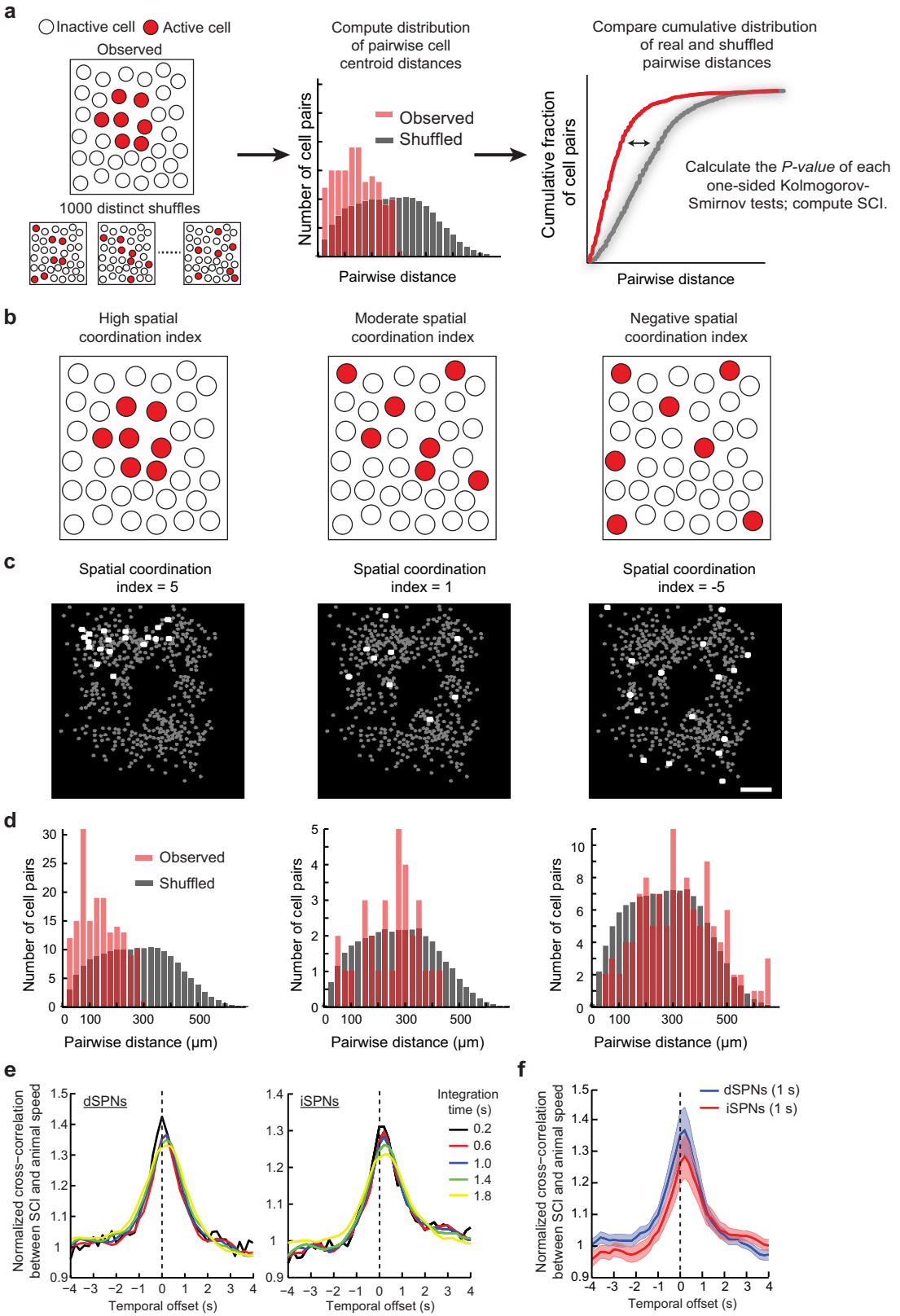


Extended Data Fig. 1 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 1 | Characterizations of striatal SPN Ca^{2+} activity patterns in live brain slices and recorded in behaving mice using a chronic microendoscopy preparation. a, We performed whole-cell patch-clamp recordings of SPNs in acute brain slices, using dual epifluorescence and infrared Nomarski (IR-DIC) microscopy to guide the recordings. We selectively recorded from fluorescent dSPNs and iSPNs in brain slices from *Drd1a*^{cre} and *Adora2a*^{cre} mice that had been injected into the striatum with AAV2/9-CAG-FLEX-GCaMP6m-WPRE.SV40. The numerical aperture used for the fluorescence recordings was matched to that used for Ca^{2+} imaging studies in freely behaving mice (Methods). **b**, Illustrative traces (top) of neural membrane potential (V_m) showing the changes that result upon stepwise injections of electrical current (bottom). The traces exhibit typical waveforms for SPNs. **c**, To visualize Ca^{2+} transients evoked by different numbers and frequencies of action potentials in iSPNs and dSPNs in brain slice recordings, we elicited spiking with 1-nA pulses (1-ms duration; 1 to 10 pulses, delivered at 5, 10, 20, 50 and 100 Hz). Top, an example of a V_m trace during 1–10 pulses injected at 10 Hz. Middle, example traces of current pulse injections delivered at 10 Hz. Bottom, a representative trace of somatic Ca^{2+} transients in response to the same 10-Hz current injections. **d**, Example Ca^{2+} activity traces for each current injection pattern tested in brain slices. As expected, electrical stimulation at higher frequencies evoked larger amplitude Ca^{2+} transients, with more sharply rising Ca^{2+} waveforms than those evoked at lower stimulation frequencies. Arrows mark current injections. **e**, **f**, Mean ± s.e.m. values for Ca^{2+} transient amplitudes (**e**) and area under the Ca^{2+} transient waveform (AUC; **f**) as a function of the number of action potentials evoked, for dSPNs and iSPNs in brain slices. Values for each cell are normalized to those evoked at the highest stimulation intensity (10 action potentials, 100 Hz). Ca^{2+} event amplitudes rose linearly with the number of action potentials ($R^2 > 0.99$ for 5–20 Hz; $R^2 > 0.97$ for all frequencies). There were no significant differences in event amplitude or AUC between dSPNs and iSPNs, at any stimulus intensity. $n = 9$ dSPNs and 8 iSPNs; two-way, mixed-design ANOVA. Exact P values can be found in the Supplementary Information for all extended data figures. **g**, **h**, Mean ± s.e.m. values for half-rise (**g**) and half-decay (**h**) times ($t_{1/2}$), measured for the Ca^{2+} events evoked in dSPNs and iSPNs in brain slices, plotted as a function of the number of action potentials. Consistent with their more prolonged stimulus durations, lower stimulation frequencies yielded greater increases in the transient rise time. Ca^{2+} transient decay times were nearly independent of action potential frequency or number. **i**, Mean ± s.e.m. values of the AUC of stimulus-evoked Ca^{2+} events in dSPNs and iSPNs in brain slices, under control conditions or perfused with SKF81297 (1 μM) or quinpirole (10 μM), respectively. Neither drug significantly affected somatic Ca^{2+} event AUCs in either SPN type, regardless of stimulus intensity. $n = 6$ dSPNs and 6 iSPNs; two-way, repeated-measures ANOVA. The individual data points are plotted alongside the mean values. **j**, Two example time-lapse fluorescence image series, acquired over the course of experiments in a *Drd1a*^{cre} mouse (top) and an *Adora2a*^{cre} mouse (bottom). Scale bars are 125 μm and apply uniformly to all panels within each row. Insets show example Ca^{2+} transients in the absence and presence of dopamine receptor agonist SKF81297 or quinpirole. Scale bars are uniformly sized for all insets. **k**, Surgical implantation of the microendoscope did not affect the ability of mice to improve their motor performance on the accelerating rotarod assay, as shown by their increased latencies to fall off the rod, which parallel the performance improvements of control mice without an implanted microendoscope. $P = 0.8$; two-way, repeated-measures ANOVA; $n = 13$ control and 20 implanted mice. Inset, Microendoscope implantation did not alter levels of spontaneous movement in an open field arena. $P = 0.7$; $n = 5$ control and 6 implanted mice; Wilcoxon rank-sum test. Data points from individual mice are shown as open circles. **l**–**n**, Box-and-whisker plots of the rates (**l**), amplitudes (**m**) and full-width at half-maximum (FWHM) durations (**n**) of Ca^{2+} events observed in live normal (untreated) mice in individual dSPNs and iSPNs, before (Pre-lesion) and after (Post 14-d) dopaminergic lesions, as the mice were resting (left) or moving (right). In resting mice, Ca^{2+} event rates decreased in dSPNs and increased in iSPNs after the lesion (**l**) as characterized in Fig. 3. When the mice were moving, Ca^{2+} event rates in iSPNs were similar before and >14 days after the lesion, whereas the rates in dSPNs were depressed after the lesion. Each box-and-whisker plot is based on $n = 3{,}325$–$3{,}703$ dSPNs or iSPNs, from 12 *Drd1a*^{cre} and 13 *Adora2a*^{cre} mice, respectively, tracked before or 14 days after the 6-OHDA lesion. Horizontal lines denote median values, boxes cover the middle two quartiles and whiskers span 1.5× the interquartile range. Data denoted as 'Pre-lesion' were taken on day −5. Data denoted as 'Post 14-d' were taken on day 14.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

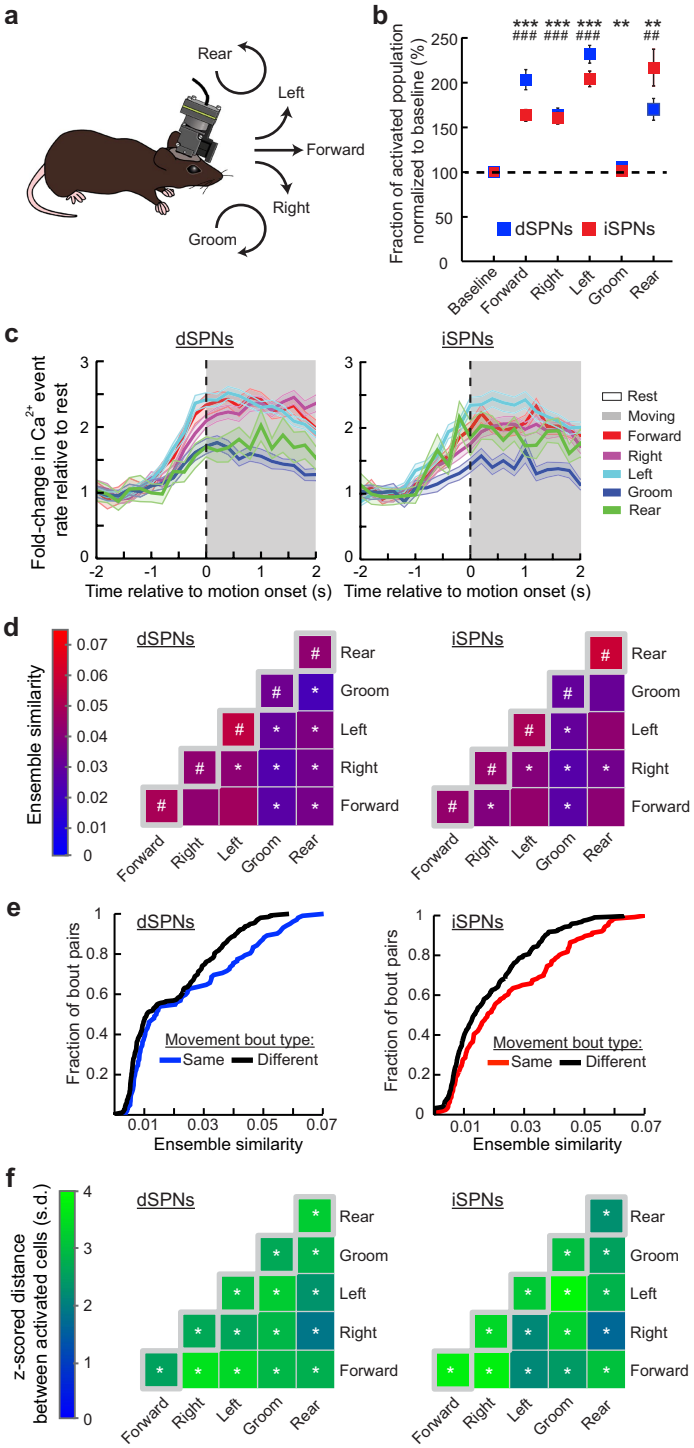


Extended Data Fig. 2 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 2 | The spatial coordination index is a metric of the extent to which the activated neurons at an individual time point are spatially clustered. a, Outline of the algorithm (Methods) used to compute the SCI at individual time points in the fluorescence Ca^{2+} movies. For each time point in a $\Delta F(t)/F_0$ movie, we created an image of all cells that were coactive within 1 s of each other. A schematic of an example image (left) has a mixture of active (red) and inactive (white) cells. We computed the set of distances between all pairs of active cells in the image (middle). We then compared these values to the set of pairwise distances between active neurons in shuffled versions of the data from the same image, in which the identities of all the individual cells were randomly permuted while maintaining the same set of cell centroid locations. The shuffled dataset as a whole comprised 1,000 different random shuffles. We compared the cumulative probability distribution function (CDF) of all pairwise distances between active cells in the real image, to that for the shuffled data (right). We then tested two hypotheses, that the activated cells in the real image were either more spatially clustered than expected from a random pattern of activation, or less spatially clustered than expected from random activation. To test these two hypotheses, we performed a pair of one-sided Kolmogorov–Smirnov tests, comparing the real CDF values to those of the shuffled data. We determined the SCI by selecting the smaller of the two *P* values from these two tests, taking its logarithm, and assigning the sign of the index according to whether the CDF of the real data showed greater (positive sign) or less (negative sign) spatial clustering than the CDF of the shuffled data. **b**, Depictions of three images with different SCI values. The left image has an SCI >0, due to its many neighbouring coactive cells. The middle image has an SCI near zero, implying the activation pattern is consistent with that of a random distribution. The right frame has an SCI <0, indicating the active cells are further apart than generally expected of random activity. **c**, Examples of three image frames of different SCI values, from an actual Ca^{2+} movie acquired in a *Drd1a*^{cre} mouse. Active cells are shown enlarged and shaded white, whereas inactive cells are shaded grey. On the left is an image with multiple neighbouring coactive neurons and a high SCI. The middle image has a modest but positive SCI. The right image has a negative SCI, indicating that the activated cells are further apart than expected given a random pattern of activation. **d**, Distributions of pairwise distances between all pairs of active dSPNs for the three corresponding image frames shown in c. The distributions for the real data are shown in red, whereas those for the shuffled datasets are shown in grey. **e**, Mean time-dependent cross-correlation functions between mouse speed and the SCI of Ca^{2+} activity in dSPNs (left) and iSPNs (right). Cross-correlation values are shown normalized to their values for abscissa values of ±4 s (Methods). We computed the SCI from Ca^{2+} event traces to which we applied forward-smoothing, using one of five different filter values (coloured curves; Methods). Error bars are omitted for clarity in **e**, but are comparable to those shown in **f**. The centre-of-mass of the cross-correlation functions occurred at positive time lags (around 12–60 ms), indicating that rises in spatially coordinated Ca^{2+} activity preceded increases in mouse speed ($P < 0.06$ for both genotypes and all filter time constants; Wilcoxon signed-rank test), but there were no significant differences across the two mouse lines ($P > 0.05$ for all filter time constants; Wilcoxon rank-sum test; $n = 37$ cross-correlation functions from 16 *Drd1a*^{cre} mice and $n = 52$ from 21 *Adora2a*^{cre} mice). **f**, Mean cross-correlation functions between mouse speed and the SCI of Ca^{2+} activity in dSPNs and iSPNs computed from Ca^{2+} event rasters forward-smoothed with a 1,000-ms time constant. Shading indicates s.e.m. **e**, **f**, Values taken from $n = 16$ *Drd1a*^{cre} mice and $n = 21$ *Adora2a*^{cre} mice, aggregated over the 1-h recordings on day −5 and the 30-min recordings performed on days −4 to −1 after saline vehicle injection but before drug administration.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

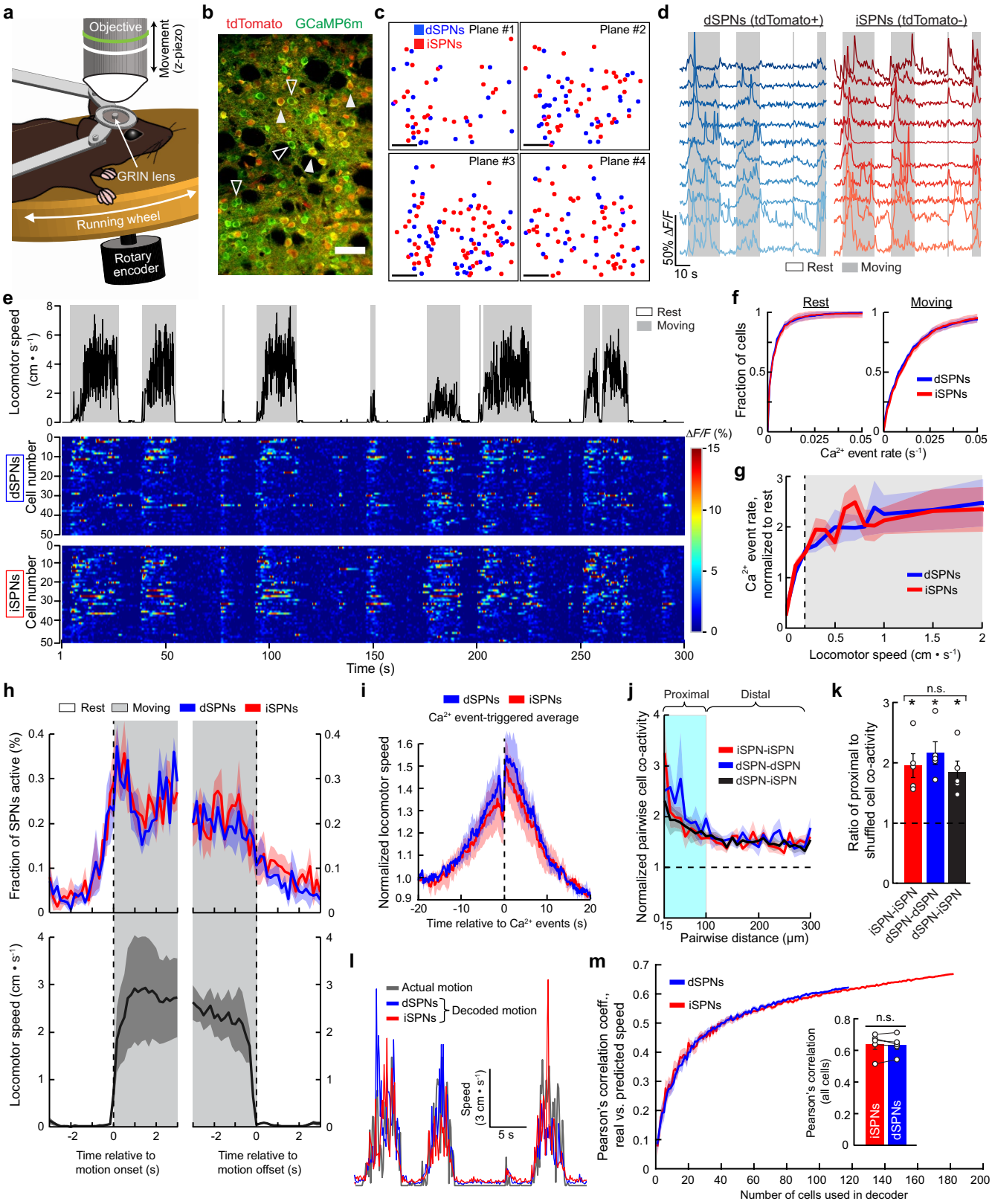


Extended Data Fig. 3 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 3 | Different types of movements recruit distinct but overlapping populations of spatially proximal dSPNs and iSPNs. **a**, To study and compare movements of different types made by freely moving mice, we used custom software to extract from the behavioural videos the intervals from −4 to 4 s surrounding the onset of each movement bout. Using this software, we manually labelled each bout as an instance of forward locomotion, a left or right turn, grooming or upward rearing. If the mouse made multiple types of movement within an individual bout, we labelled the bout according to the first movement type exhibited, as only the interval from −1 to 2 s relative to motion onset was used in subsequent analyses of the accompanying neural Ca^{2+} activity. **b**, The fraction (mean ± s.e.m.) of SPNs that exhibited Ca^{2+} activity, relative to the baseline periods immediately preceding each movement type. Relative to baseline (dashed line), there was a significant increase in the fraction of dSPNs and iSPNs activated for all movement types except grooming. $^{**}P < 5 \times 10^{-3}$ and $^{***}P < 5 \times 10^{-8}$ for dSPNs; $^{\#\#}P < 10^{-4}$ and $^{\#\#\#}P < 10^{-10}$ for iSPNs; Wilcoxon signed-rank test. Data in **b** and **c** are from $n = 492$ forward movements, 657 right turns, 810 left turns, 732 grooming and 204 rearing bouts from 17 *Drd1a*^{cre} mice, and $n = 790$ forward, 785 right turns, 1,015 left turns, 792 grooming and 164 rearing bouts from 21 *Adora2a*^{cre} mice. **c**, Rates (mean ± s.e.m.) of Ca^{2+} events in dSPNs and iSPNs, plotted as a function of time relative to the onsets of different types of movements. Event rates are shown normalized to the values of −2 to −1 s before motion onset and rose significantly above baseline values during all types of movement in both cell types. $P < 10^{-6}$ for both cell types and all movement types; Wilcoxon signed-rank test. **d**, Mean values of the neural ensemble similarity computed for the sets of dSPNs and iSPNs that were active during pairs of bouts of either the same (on-diagonal) or different (off-diagonal) types of movements (Methods). For all movement types, the similarities of the cell ensembles that were active on different bouts of the same movement type were significantly greater than those of the ensembles that were active during the baseline periods before each bout. $^{\#}P < 0.05$; Kolmogorov–Smirnov test, corrected for multiple comparisons using a Benjamini–Hochberg procedure with a false-discovery rate of 0.05. Off-diagonal asterisks indicate that the neural ensembles that were active during bouts of two different movement types were significantly less similar to each other than the ensembles activated on different bouts of the same movement type, for both of the two movement types under consideration. $^{*}P < 0.05$; Kolmogorov–Smirnov test, corrected for multiple comparisons using a Benjamini–Hochberg procedure with a false-discovery rate of 0.05. **e**, The cumulative distribution functions show the range of ensemble similarity values for the dSPN and iSPN ensembles that were activated on two bouts of the same movement type or on two bouts of different movement types (as described in **d**). For both iSPNs and dSPNs, ensemble similarity values were significantly lower for two bouts of different movement types than for two different bouts of the same movement type. $P < 0.01$; Kolmogorov–Smirnov test. **f**, Mean pairwise distances between the individual dSPNs or iSPNs activated on bouts of two different movement types. We normalized the values by comparing the mean weighted distance between the cells that were active during the two movement types to the same quantity determined under the null hypothesis that the spatial probability distributions of active cells on the two movement types were the same. For the latter determination, we created shuffled datasets in which we randomly permuted the firing rate of each cell between the two movement types, and we averaged the results over twenty-five different shuffled datasets. We normalized each distance value by taking the actual mean value, subtracting the mean value determined under the null hypothesis, and then dividing this difference by the standard deviation of the distance across the twenty-five shuffled datasets (Methods). Asterisks indicate that the mean pairwise distance was significantly greater than that expected by chance, indicating that the two movement types activated spatially distinguishable cell ensembles. $^{*}P < 0.05$; Wilcoxon signed-rank test, corrected for multiple comparisons using a Benjamini–Hochberg procedure with a false-discovery rate of 0.05. Plots in **d**–**f** are based on $n = 102$ and $n = 126$ comparisons between bouts of the same movement type, and 255 and 315 comparisons of bouts of different movement types, in *Drd1a*^{cre} and *Adora2a*^{cre} mice, respectively. Values are from $n = 17$ *Drd1a*^{cre} mice and $n = 21$ *Adora2a*^{cre} mice, aggregated over the 1-h recordings on day −5 and the 30-min recordings performed on days −4 to −1 after saline vehicle injection but before drug administration.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

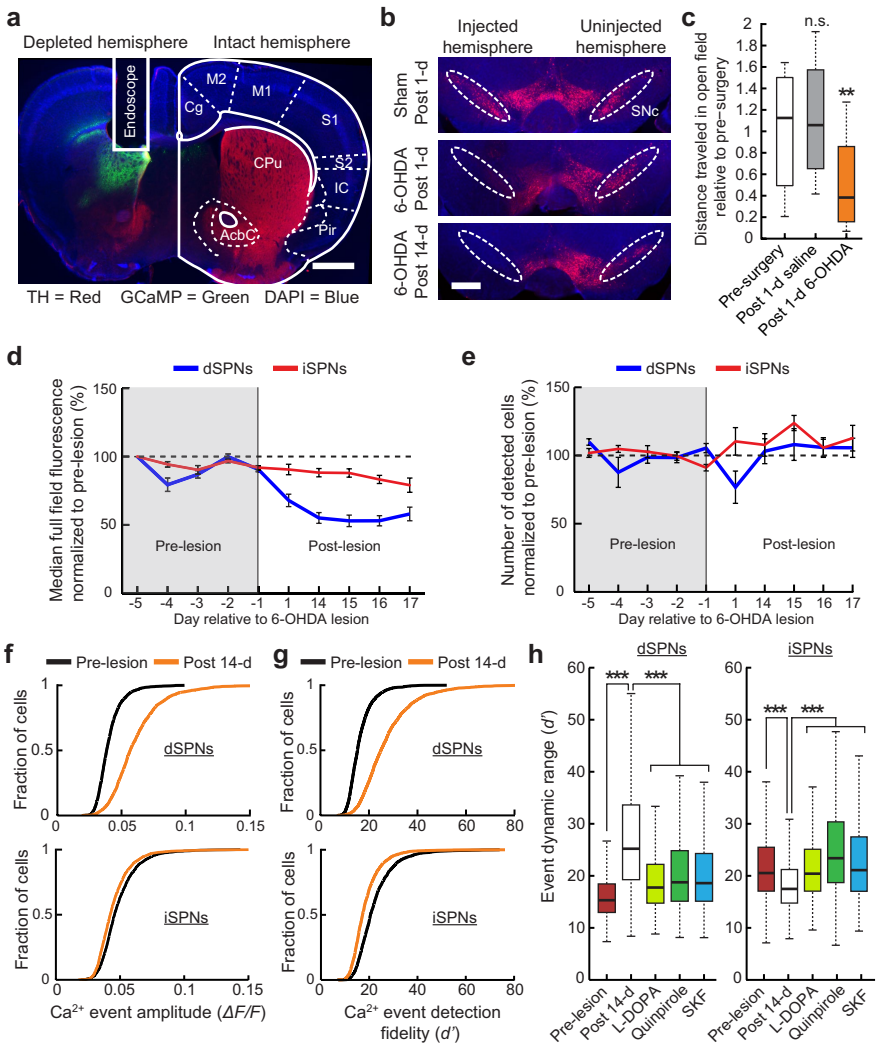


Extended Data Fig. 4 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 4 | Simultaneous recordings of dSPNs and iSPNs show that the two cell types encode movement with indistinguishable spatiotemporal patterns of activation. **a**, To image dSPN and iSPN Ca^{2+} activity concurrently, we prepared mice that expressed GCaMP6m in both cell types but tdTomato only in dSPNs (Methods). Mice were head-fixed on a running wheel beneath the objective lens of a two-photon microscope. A piezoelectric actuator moved the axial position of the objective to allow volumetric imaging across four planes at different depths of the tissue (15-μm axial spacing). The mice were free to run or rest on the wheel (Supplementary Video 3). We tracked the motion of the wheel using a rotary encoder (500 encoder pulses per revolution) that provided a read-out of the instantaneous locomotor speed. We computed the mean speed of the mouse at a time-resolution matching that of the two-photon volume acquisition rate (6 Hz, or 166 ms per time bin). We identified periods of movement by marking all time bins in the mean speed trace with values $>0.2\,cm\,s^{-1}$. To identify instances of motion onset, we selected all time bins for which speeds were $<0.2\,cm\,s^{-1}$ for at least 1 s in the immediately prior time bins, and $>0.2\,cm\,s^{-1}$ for at least 1 s in the immediately subsequent time bins. To identify instances of motion offset, we used the opposite criterion. **b**, Histological section of dorsomedial striatum expressing tdTomato (red) in *Drd1a*^{cre} positive cells and immunostained for GFP (green) to visualize GCaMP6m expression. Closed arrowheads point to three example dSPNs that expressed both GCaMP6m and tdTomato. Open arrowheads point to three putative iSPNs that expressed GCaMP6m but not tdTomato. Scale bar, 50 μm. **c**, Representative cell maps from each of the four imaging planes in an example mouse with 118 detected dSPNs (blue) and 183 detected iSPNs (red). Scale bars, 100 μm. **d**, Representative traces of Ca^{2+} activity from 10 dSPNs (blue) and 10 iSPNs (red) from the same mouse as in **c**. Grey shading here and in **e**, **g**, **h** denotes periods classified as movement on the running wheel. **e**, Locomotor speed on the running wheel (top) and Ca^{2+} activity traces of individual dSPNs (middle) and iSPNs (bottom), during part of an imaging session in an example mouse. Note the clear correlation between locomotion and Ca^{2+} activity in both cell types. **f**, Mean cumulative distribution functions of Ca^{2+} event rates in dSPNs ($n = 699$) and iSPNs ($n = 1020$) were nearly identical, during periods of rest (left) and running (right). **g**, Mean Ca^{2+} event rates in dSPNs and iSPNs as a function of mouse locomotor speed. Events rates are shown normalized to their mean levels when the mice were resting (speed $<0.2\,cm\,s^{-1}$). Grey shading denotes speeds at which we classified the mouse as moving ($>0.2\,cm\,s^{-1}$). **h**, Mean locomotor speed (bottom) and the fraction of SPNs that are activated (top), plotted as a function of time relative to motion onset (left) and offset (right). We determined the onset time of neural activity as the time at which the mean percentage of active cells was ≥ 3 s.d. of the percentage of active cells during the baseline periods of −3 to −1 s relative to motion onset. Using this criterion, dSPNs and iSPNs respectively activated -666 ± 97 ms and -733 ± 51 ms before motion onset (mean ± s.e.m; $n = 5$ mice; Methods), which were statistically indistinguishable. $P = 0.9$; Wilcoxon rank-sum test. **i**, Mean locomotor speed as a function of time relative to the occurrences of Ca^{2+} events in dSPNs and iSPNs. For each cell, we normalized these traces to the mean speed 20 s before Ca^{2+} excitation, then averaged the traces across all cells of each type and all mice. **j**, Mean pairwise coactivity (Methods) for dSPN–dSPN, iSPN–iSPN and dSPN–iSPN cell pairs, computed as a function of the distances between the pairs of cells, normalized to coactivity values in temporally shuffled datasets (1,000 distinct shuffles) in which time-correlated activity patterns were scrambled. Cyan shading indicates proximal (20–100 μm) cell pairs. Data are from periods of movement on the running wheel. **k**, Pairwise coactivity values (mean ± s.e.m.) were significantly greater for proximal cell pairs than those in temporally shuffled datasets (*$P < 0.05$; Wilcoxon rank-sum test; $n = 5$ mice), and did not depend on the SPN types. n.s. denotes $P > 0.05$; Wilcoxon rank-sum test. Data points from individual mice are shown as open circles. **l**, We created GLMs to make time-dependent predictions of mouse locomotor speed based on the $\Delta F(t)/F_0$ activity traces of the cells determined by two-photon Ca^{2+} imaging. We used the GLM libraries in MATLAB (Mathworks), using a Gaussian noise model and taking the identity function as the linking function. We used 70% of the time bins in the set of $\Delta F(t)/F_0$ traces for training the GLM, and 30% for testing it. To study the accuracy of the speed predictions as a function of the number of cells (n) included in the GLM, for each value of n, we randomly chose n cells from the total available, constructed the GLM and computed the Pearson's correlation coefficient between the actual and predicted speed traces. For each n value, we repeated this procedure with 10 different randomly chosen subsets of cells and then computed the mean correlation coefficient of the real and predicted speeds across the 10 different sub-samplings. In an example mouse, traces of the actual running speed (grey) were well fit by models based on the activity of either dSPNs (blue) or iSPNs (red). **m**, Mean Pearson correlation coefficients for an example mouse between the actual and predicted running speeds, for GLMs based on either dSPNs or iSPNs (10 sets of randomly chosen cells for each abscissa value, out of 301 total cells). Inset shows the correlation coefficients (mean ± s.e.m.) from $n = 5$ mice, computed using either 115 dPSNs or 115 iSPNs from each mouse. Data points from individual mice are shown as open circles and were statistically indistinguishable across the two cell types. $P = 0.13$; Wilcoxon signed-rank test. **f**–**j**, Colour shading denotes s.e.m. for $n = 5$ *Drd1a*^{cre} × Ai14 mice.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

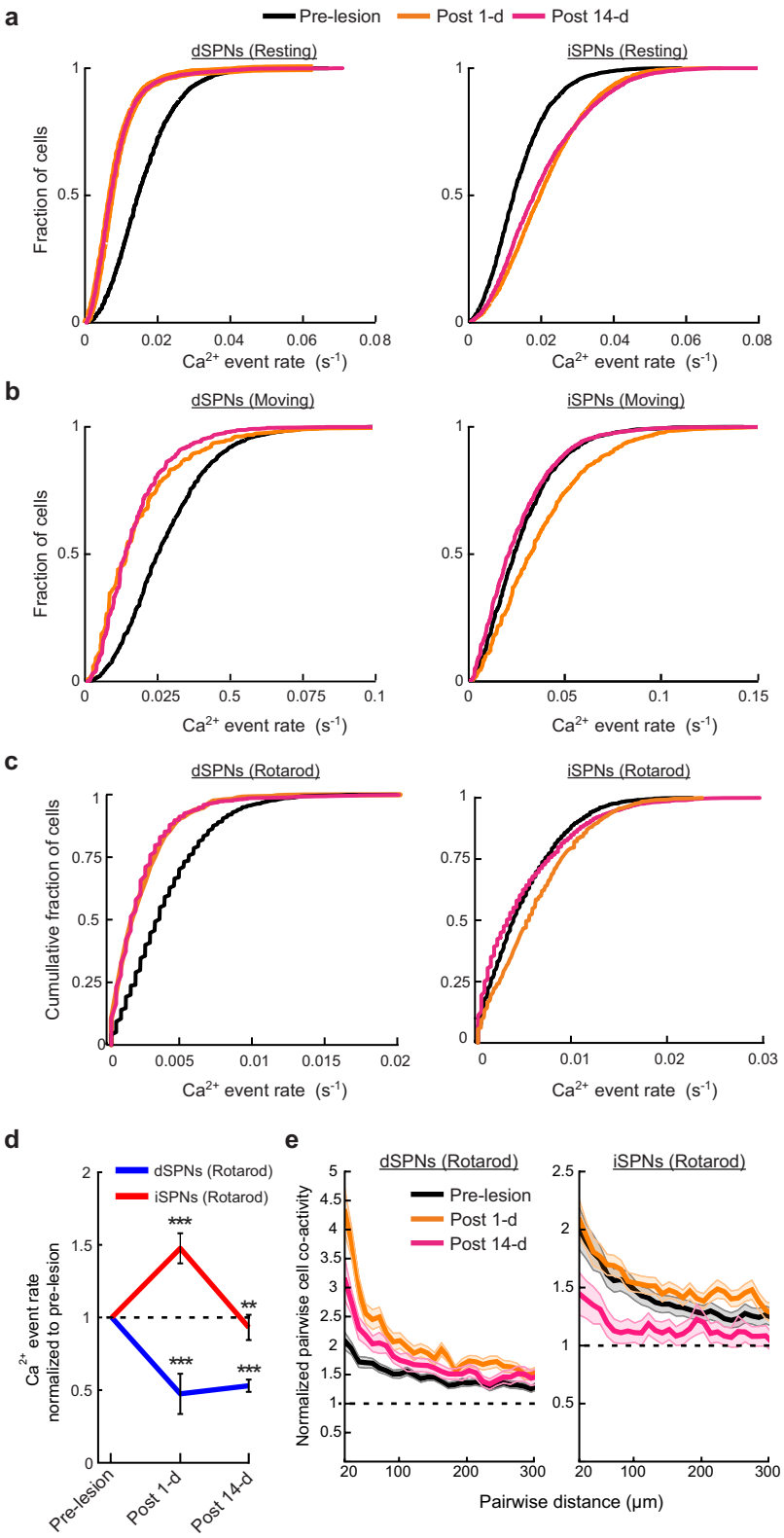


Extended Data Fig. 5 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 5 | 6-OHDA injections rapidly ablate SNc dopamine neurons and disrupt motor behaviour without impairing the fidelity of Ca^{2+} event detection. a, Example coronal brain section from an experimental mouse, immunostained for GCaMP6m (anti-GFP, green) and tyrosine hydroxylase (anti-TH, red). DAPI was used to stain the nuclei (blue). White lines demarcate the position of the implanted microendoscope, and the boundaries of nearby brain areas. AcbC, accumbens core; Cg, cingulate cortex; CPu, caudate/putamen; IC, insular cortex; M1 and M2, motor cortices; Pir, piriform cortex; S1 and S2, somatosensory cortices. Labels are adapted from an anatomical atlas of the mouse brain[38]. Scale bar, 1 mm. **b**, Representative midbrain coronal sections acquired 1 day after unilateral infusions of saline (top) or 6-OHDA (middle), or >14 days after 6-OHDA infusion (bottom), and then immunostained for tyrosine hydroxylase (red). DAPI was used to stain the nuclei (blue). Dopamine cell bodies are absent in the SNc of mice that received 6-OHDA, at both 1 and >14 days after 6-OHDA infusion. Scale bar, 500 μm. **c**, 24 h after infusions into SNc of 6-OHDA but not of saline, mice exhibited disrupted patterns of spontaneous locomotion in an open field arena. $P = 4 \times 10^{-3}$ for 6-OHDA and $P = 0.8$ for saline; $n = 7$ saline-treated and 14 6-OHDA-treated mice; Wilcoxon signed-rank test for comparisons to the pre-lesion behaviour of each mice. **d**, The median fluorescence intensity across the entire imaging field (normalized to pre-lesion values on day −5 for each mouse and then averaged across mice) decreased significantly after 6-OHDA lesions in mice of both genotypes. $P < 10^{-7}$ comparing median fluorescence intensities (normalized to pre-lesion means) averaged across 5 days before versus 5 days after lesion; Wilcoxon rank-sum test. However, fluorescence intensities stabilized 15–17 days after the 6-OHDA lesion, and there were no further significant changes over time in either mouse line. $P > 0.05$; $n = 5$ *Drd1a*^{cre} and 7 *Adora2a*^{cre} mice; Spearman correlation. Error bars indicate s.e.m. for 5 *Drd1a*^{cre} and 7 *Adora2a*^{cre} mice. **e**, The number of active SPNs (mean ± s.e.m.) detected in total on each day of the study, normalized to the mean value (dashed horizontal line) detected before 6-OHDA infusion (days shaded grey). On days on which mice received drug treatments (Fig. 1a), the data shown here are from the initial portions of the recording sessions before drug administration. The number of active cells was stable across the study, except for a single pairwise difference in the number of active iSPNs. Friedman ANOVA; $n = 5$ *Drd1a*^{cre} and $n = 7$ *Adora2a*^{cre} mice; $P > 0.05$ for *Drd1a*^{cre} and $P = 0.01$ for *Adora2a*^{cre}; $P = 5 \times 10^{-4}$ for post hoc test comparing the number of iSPNs detected 1 day before the lesion and 15 days after the lesion; Fisher's least significant difference test with a Holm–Bonferroni correction for multiple comparisons. **f**, Cumulative distributions of peak $\Delta F/F$ values for Ca^{2+} events from individual dSPNs (top) and iSPNs (bottom), before and after 6-OHDA lesion. After the lesion, Ca^{2+} event amplitudes were significantly greater in dSPNs (3.7 ± 0.02% (pre-lesion) versus 6.0 ± 0.03% (14 days after)), but smaller in iSPNs (4.4 ± 0.02% (pre-lesion) versus 4.0 ± 0.02% (14 days after)). Data are mean ± s.e.m.; $P < 10^{-10}$ for both SPN types; Wilcoxon rank-sum test. Data in **f** and **g** are from $n = 3{,}332$–$3{,}734$ dSPNs or iSPNs, from before (day −5) or 14 days (day 14) after 6-OHDA lesion, in $n = 12$ *Drd1a*^{cre} mice and 13 *Adora2a*^{cre} mice, respectively. **g**, Cumulative distributions of the signal detection fidelity (d') of Ca^{2+} events in individual dSPNs (top) and iSPNs (bottom), before and after 6-OHDA lesion. Owing to the decrease in background fluorescence intensity, Ca^{2+} events in dSPNs became easier to detect after 6-OHDA lesion, as quantified by the increase in d' values for dSPNs. $d' = 17 \pm 0.1$ (mean ± s.e.m.; pre-lesion) versus 29 ± 0.2 (14 days after). The changes in d' values for iSPNs were smaller in magnitude ($d' = 22 \pm 0.1$ (pre-lesion) versus 19 ± 0.1 (14 days after)), while keeping the optical conditions for Ca^{2+} event detection extremely favourable. Crucially, all changes in Ca^{2+} event detection fidelity values were opposite to those observed for Ca^{2+} event rates in the two cell types (Fig. 3), and thus cannot account for the event rate changes. **h**, d' values for Ca^{2+} events from individual dSPNs (left) and iSPNs (right), before and after 6-OHDA lesions, and after drug treatments in mice following 6-OHDA lesion. d' values increased after 6-OHDA lesions and decreased after the drug treatments in dSPNs. By contrast, d' values decreased after 6-OHDA lesions and increased after drug treatment in iSPNs. $^{***}P < 10^{-10}$ for all conditions in both genotypes; Wilcoxon rank-sum test; values are from $n = 1{,}770$–$2{,}027$ dSPNs from 5 *Drd1a*^{cre} mice and $n = 1{,}719$–$2{,}393$ iSPNs from 6 *Adora2a*^{cre} mice, recorded before (day −5) and after 6-OHDA lesion (day 14), or after the lesion and quinpirole (day 16), SKF81297 (day 18) and L-DOPA (day 20) treatments. As in **g**, these changes in d' have an opposite sign to the changes in Ca^{2+} event rates (Fig. 4), and thus cannot account for the latter effects. Throughout the study, d' values remained extremely high, in that a d' value >16 corresponds mathematically to a mean rate of $<10^{-10}$ errors in Ca^{2+} event detection per hour. This nearly vanishing predicted error rate is unattainable experimentally over many hours of recording but underscores the highly favourable conditions for Ca^{2+} imaging.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

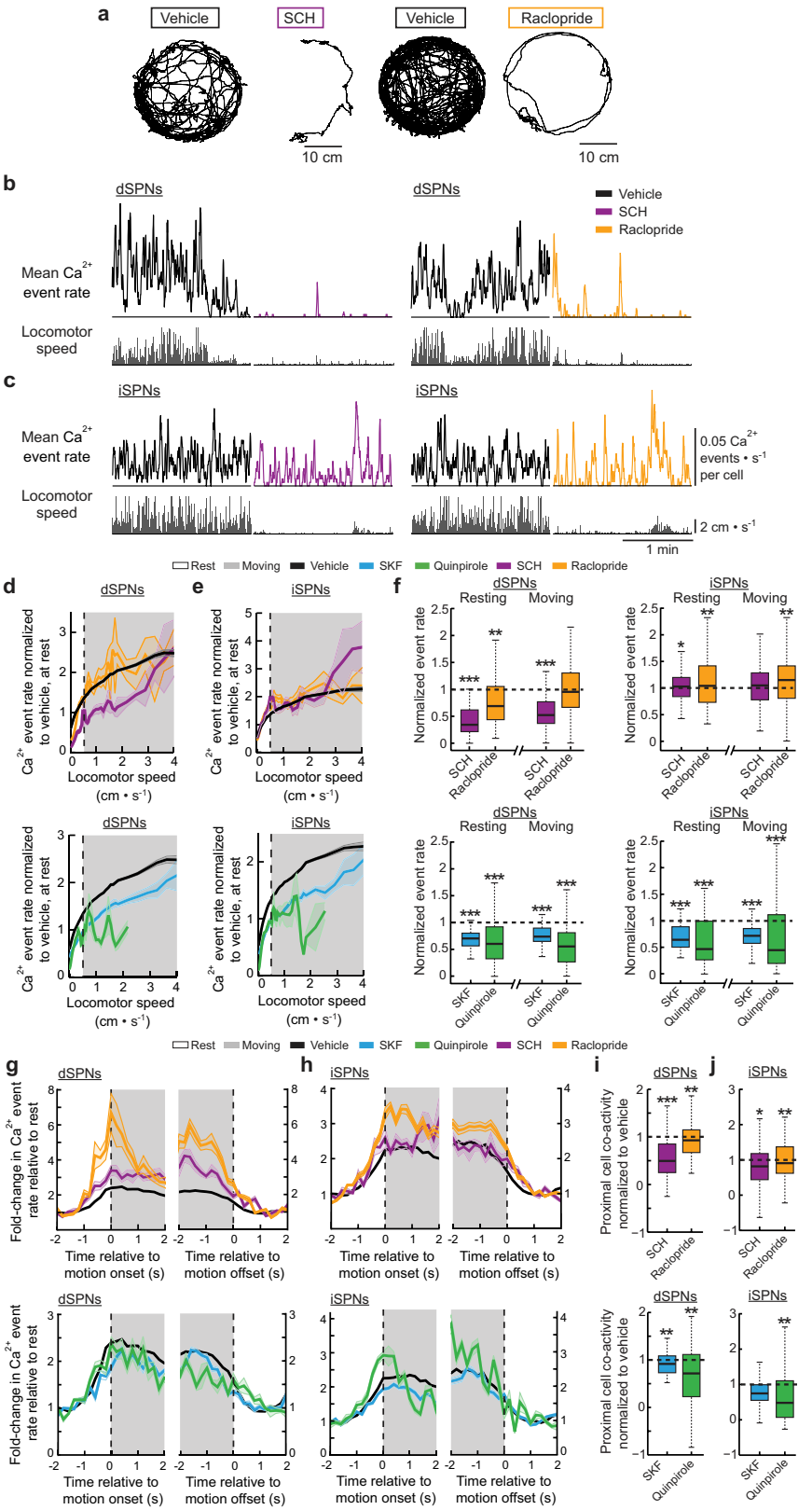


Extended Data Fig. 6 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 6 | Lesion of dopamine neurons altered Ca^{2+} event rates in dSPNs and iSPNs during spontaneous open field exploration and forced movement on the rotarod. a, **b**, Cumulative probability distributions of Ca^{2+} event rates in dSPNs (left) and iSPNs (right) while mice were resting (**a**) or in locomotion (**b**) before the lesion of dopamine neurons, 1 day after the lesion (day 1), and >14 day after the lesion (day 14). Ca^{2+} event rates in dSPNs were depressed at both time points after the lesion, during rest and locomotion. Ca^{2+} event rates in iSPNs were elevated at both time points after the lesion when mice were resting. However, when the mice were moving, iSPN activity rates were elevated at 1 day but not >14 days after the lesion, compared to values from before the lesion. The distributions of Ca^{2+} event rates in dSPNs in resting mice were nearly identical at 1 and >14 days, making it hard to visually distinguish the two plots. Data are from 12 *Drd1a*^{cre} mice and 13 *Adora2a*^{cre} mice. The *Drd1a*^{cre} mice yielded a total of 2,554–3,732 activated dSPNs during movement and rest, across the different days of the study, whereas the *Adora2a*^{cre} mice yielded 3,209–3,702 iSPNs. All comparisons within each SPN type between Ca^{2+} event rates from before the lesion to those 1 d and 14 d after the lesion were significant. $P < 10^{-3}$; Wilcoxon rank-sum test. **c**, Cumulative distribution functions of Ca^{2+} event rates in individual dSPNs and iSPNs as mice walked on the rotarod. At 1 day after dopamine depletion (day 1), Ca^{2+} event rates were reduced in dSPNs and increased in iSPNs. At >14 days after dopamine depletion (day 14), Ca^{2+} event rates were still reduced in dSPNs but in iSPNs were at or even slightly below normal values. Data pre-lesion and >14 days after dopamine depletion are based on 5 *Drd1a*^{cre} mice and 7 *Adora2a*^{cre} mice; however, 1 day after dopamine depletion only 2 *Drd1a*^{cre} mice and 3 *Adora2a*^{cre} mice could perform the rotarod assay without falling off, precluding determinations of Ca^{2+} event rates from the other mice on that day. All comparisons within each SPN type between Ca^{2+} event rates from before the lesion to those 1 and >14 days after the lesion were highly significant ($P < 10^{-10}$; Wilcoxon rank-sum test), except for the comparison of iSPN activity rates from before the lesion to those >14 days after ($P = 5 \times 10^{-3}$; Wilcoxon rank-sum test). **d**, Rates of Ca^{2+} events (mean ± s.e.m.) as mice walked on the rotarod, normalized to the corresponding values from before dopamine depletion. dSPN activation rates were substantially reduced at both time points after the lesion. iSPN event rates were substantially increased 1 day after the lesion (day 1), but had returned to near baseline levels by 14 days after the lesion (day 14), mirroring the effects of dopamine depletion on dSPN and iSPN activity during spontaneous movement (see Fig. 3f). $^{**}P < 5 \times 10^{-3}$ and $^{***}P < 5 \times 10^{-7}$; Wilcoxon rank-sum test. Owing to the large numbers of Ca^{2+} events across thousands of SPNs that contributed to these calculations, the slight depression of Ca^{2+} event rates in iSPNs at 14 days after the lesion was statistically significant. **e**, The extent of pairwise coactivation of dSPNs (left) and iSPNs (right) during the rotarod assay, as a function of the pairwise distance between cells, before dopamine depletion, at 1 and >14 days after dopamine depletion. Values are normalized to those attained from shuffled datasets in which the times of the Ca^{2+} events of each cell were randomized (Methods). The spatial clustering of activity in proximal iSPN pairs (20–100 μm apart) was significantly reduced 14 days after the lesion, relative to values from beforehand. $P < 10^{-16}$; Wilcoxon rank-sum test; $n = 7$ *Adora2a*^{cre} mice. Shaded areas denote s.e.m. Data and statistical tests in **c**–**e** are from 1,930, 712 and 1,672 dSPNs and 1,731, 853 and 1,607 iSPNs that were active during the rotarod assay pre-lesion, 1 (day 1), and 14 (day 14) days after the lesion, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

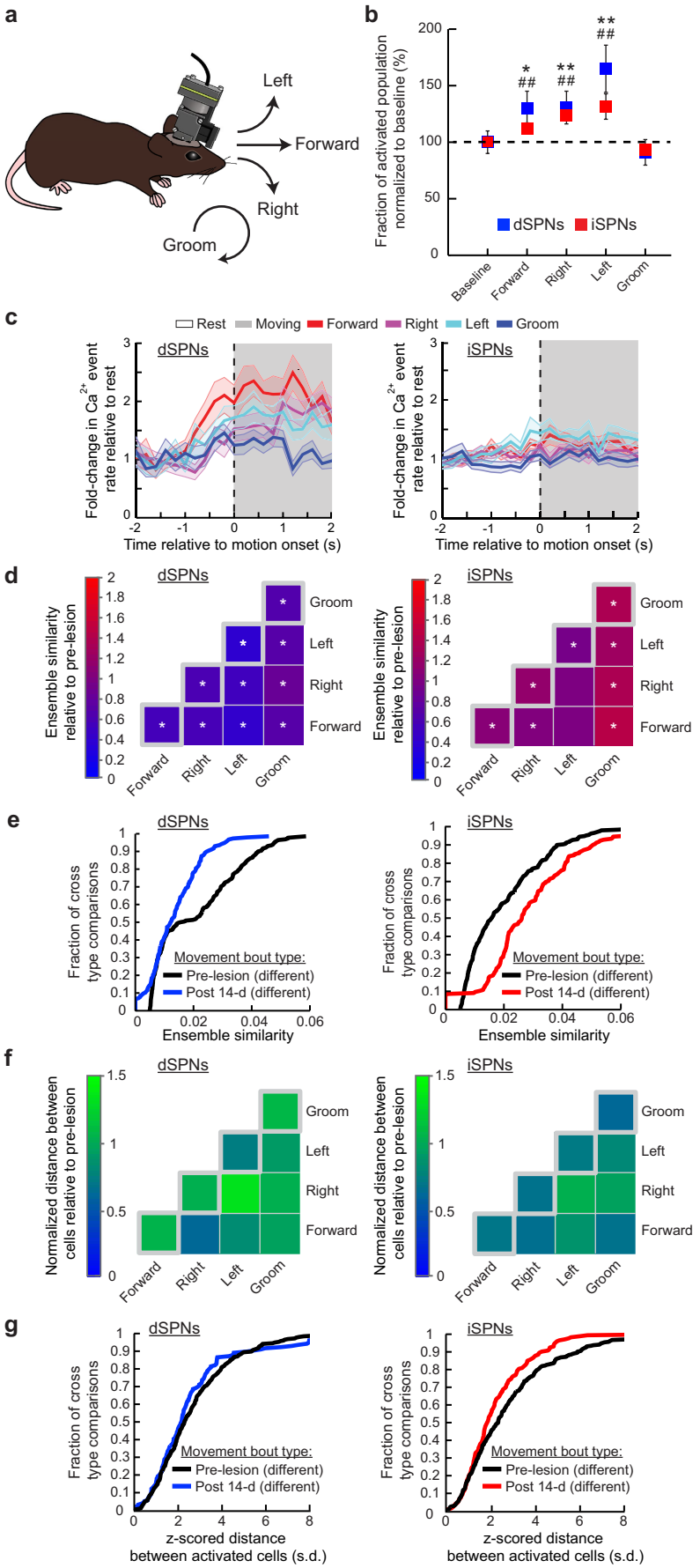


Extended Data Fig. 7 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 7 | In healthy mice, D1R or D2R antagonists affect SPN activity similarly to an acute (but not chronic) loss of dopamine neurons, whereas dopamine receptor agonists have divergent effects from those observed in dopamine-depleted mice. a–c, Example locomotor trajectories (15-min duration) of mice moving freely within a circular arena (**a**) and paired example traces of mouse locomotor speed and the mean rate of Ca^{2+} events in dSPNs (**b**) and iSPNs (**c**) after administration of saline vehicle or a selective antagonist of D1Rs (SCH23390, 0.2 mg kg^{-1}) or D2Rs (raclopride, 1 mg kg^{-1}). Both drugs reduced locomotion relative to vehicle. $P = 10^{-4}$; Wilcoxon signed-rank test; $n = 18$ mice. **d**, **e**, Mean Ca^{2+} event rates in dSPNs (**d**) and iSPNs (**e**) as a function of locomotor speed, following administration of saline vehicle, a D1R- or D2R-selective antagonist (0.2 mg kg^{-1} SCH23390 or 1 mg kg^{-1} raclopride, respectively) (top) or a D1R- or D2R-selective agonist (1 mg kg^{-1} SKF81297 or 1 mg kg^{-1} quinpirole, respectively) (bottom). Rates are normalized to cell population means when mice were at rest (<0.5 cm s^{-1}) after vehicle injection (Methods). Colour shading surrounding each curve denotes s.e.m. **f**, SPN Ca^{2+} event rates during rest and movement after treatment with a dopamine-receptor antagonist or agonist. The D1R antagonist (SCH23390) reduced dSPN but not iSPN activity. The D2R antagonist (raclopride) increased iSPN but not dSPN activity. Both agonists (quinpirole, SKF81297) reduced activity in both SPN types during rest and movement. $*P < 0.05$, $**P < 5 \times 10^{-3}$, $***P < 5 \times 10^{-7}$; Wilcoxon signed-rank test comparing all drug conditions to saline vehicle injection; $n = 14$ speed bins per mouse in the resting state (<0.5 cm s^{-1}) and 24 speed bins per mouse in the moving state (>0.5 cm s^{-1}); $n = 7$ *Drd1a*^{cre} and 11 *Adora2a*^{cre} mice. **g**, **h**, Mean Ca^{2+} event rates in dSPNs (**g**) and iSPNs (**h**) relative to motion onset (left) and offset (right) following administration of saline vehicle, SCH23390 or raclopride (top), or SKF81297 or quinpirole (bottom). Traces are normalized to cell population means averaged over the time period from −2 to −1 s preceding motion onset. Raclopride increased dSPN ($P = 6 \times 10^{-11}$) and iSPN ($P = 2 \times 10^{-9}$) activation at motion onset. Wilcoxon rank-sum test; comparing 11 time bins for the interval of 0–2 s after motion onset to the baseline periods in each of the 7 *Drd1a*^{cre} and 11 *Adora2a*^{cre} mice. SCH23390 slightly altered dSPN ($P = 0.02$) but not iSPN activation ($P = 0.4$) at motion onset. In *Drd1a*^{cre} mice, there were $n = 1{,}917$ (vehicle), 134 (SCH23390) and 129 (raclopride) instances of motion onset, whereas in *Adora2a*^{cre} mice there were 2,798 (vehicle), 155 (SCH23390) and 233 (raclopride) such instances. Colour shading surrounding each curve denotes s.e.m. **i**, **j**, Coactivity of proximal (20–100 μm apart) dSPN (**i**) or iSPN (**j**) cell pairs during periods of movement, following administration of saline vehicle, SCH23390 or raclopride (top) or saline, SKF81297 or quinpirole (bottom). Coactivity values are plotted after subtraction of the coactivity values determined for temporally shuffled datasets under the same drug treatment conditions, and then normalized to the values attained for saline administration. Comparing measured values for each drug to those attained with saline; $*P < 0.05$, $**P < 5 \times 10^{-3}$ and $***P < 5 \times 10^{-7}$; Wilcoxon signed-rank test; 8 spatial bins of cell–cell separation for each of the $n = 7$ *Drd1a*^{cre} and $n = 11$ *Adora2a*^{cre} mice. **d**–**j**, Data are based on the same 7 *Drd1a*^{cre} and 11 *Adora2a*^{cre} mice. Figure 1a shows the schedule of drug treatments. Data acquired in the 30-min Ca^{2+} imaging sessions after drug administration have been normalized for each mouse to the values determined on the same day during the 30-min Ca^{2+} imaging sessions occurring after saline vehicle administration but before drug treatment.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

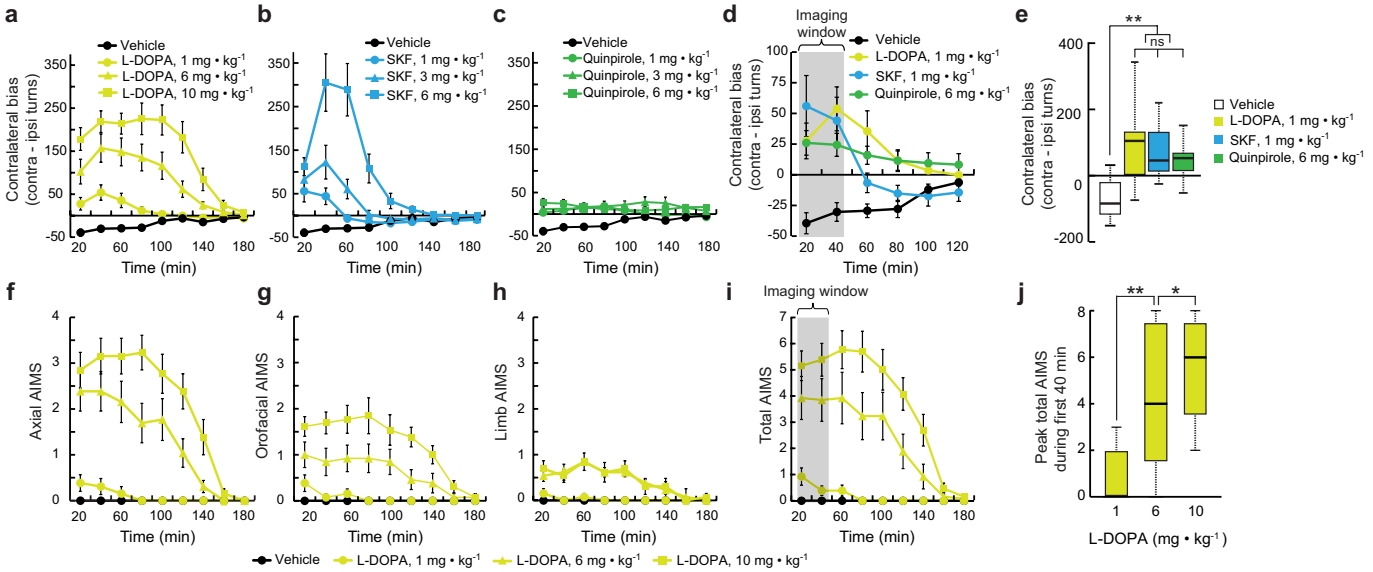


Extended Data Fig. 8 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

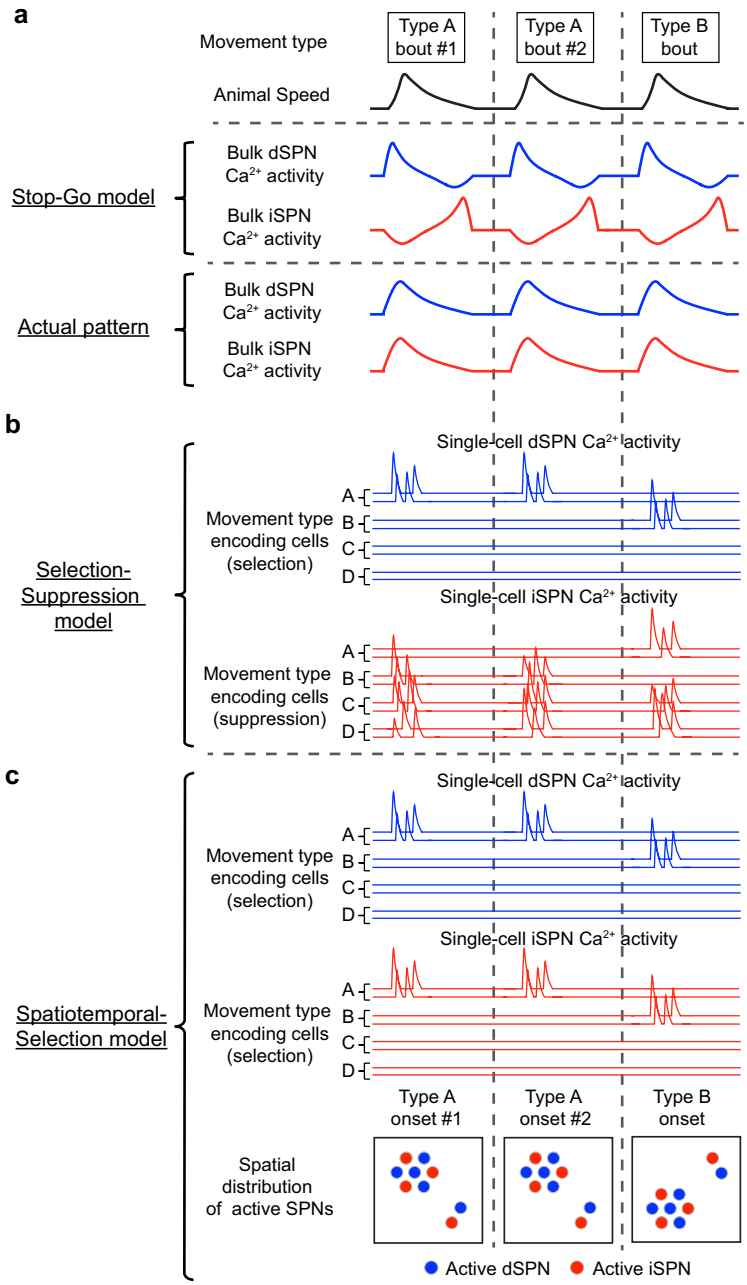
Extended Data Fig. 8 | Dopamine depletion reduces the movement-type specificity of iSPN activity patterns. a, Following dopamine depletion, we classified mouse movement bouts into different types, as in Extended Data Fig. 3. After 6-OHDA lesion, there were insufficient bouts of upward rearing to allow statistical analyses. **b**, Percentages of SPNs (mean ± s.e.m.) that exhibited Ca^{2+} activity, relative to the baseline periods (dashed line) immediately before each movement type, >14 days after 6-OHDA lesion. Compared to baseline periods, there was a significant increase in dSPN and iSPN activity for all movement types, except during grooming. $*P < 0.05$ and $**P < 5 \times 10^{-3}$ for dSPNs; $\#P < 0.05$ and $\#\#P < 5 \times 10^{-3}$ for iSPNs; Wilcoxon signed-rank test. Data in **b** and **c** are based on 206 forward movements, 200 right turns, 290 left turns and 314 grooming bouts in 13 *Drd1a*^{cre} mice and 331 forward movements, 295 right turns, 194 left turns and 339 grooming bouts in 21 *Adora2a*^{cre} mice. **c**, Rates of Ca^{2+} events (mean ± s.e.m.) in dSPNs and iSPNs as a function of time relative to the onsets of different types of movements, >14 days after 6-OHDA lesion. Event rates are shown normalized to the values from −2 to −1 s before motion onset. As found when all movements were grouped together (Fig. 3g), the rates of iSPNs hardly increase at motion onset in parkinsonian mice. **d**, Mean values of the neural ensemble similarity computed for the sets of dSPNs and iSPNs that were active during pairs of bouts of either the same (on-diagonal) or different (off-diagonal) types of movements, in freely behaving mice >14 days after 6-OHDA lesion (Methods). For each pair of movement types, values are shown normalized to the corresponding value found in healthy mice before 6-OHDA lesion. Asterisks indicate a significant difference between the pre- and post-lesion values. $*P < 0.05$; Wilcoxon rank-sum test, corrected for multiple comparisons using a Benjamini–Hochberg procedure with a false-discovery rate of 0.05. **e**, Cumulative distribution functions showing the range of ensemble similarity values for the sets of dSPNs and iSPNs that activated on two bouts of different movement types. The similarity of the dSPN ensembles activated on different movement types decreased significantly after 6-OHDA, indicating that the representations of different movements became more distinct. $P < 0.05$; Kolmogorov–Smirnov test. By comparison, the iSPN ensembles activated on different types of movements became more similar after 6-OHDA lesion, consistent with a reduction in the selectivity of iSPN movement encoding. $P < 0.05$; Kolmogorov–Smirnov test. **f**, Mean pairwise distances between the individual dSPNs or iSPNs activated on bouts of two different movement types, in freely behaving mice >14 days after 6-OHDA lesion. We expressed the values as *z* scores, as described in Extended Data Fig. 3f. For each pair of movement types, we then normalized each *z* score by the corresponding value found in healthy mice before 6-OHDA lesion. **g**, Cumulative distribution functions showing the normalized distances between dSPNs (left) and iSPNs (right) (computed as in **f**) before and after 6-OHDA lesion, for bouts of different movement types. The mean pairwise distances between iSPNs, but not dSPNs, that activated on bouts of different movement types decreased significantly after 6-OHDA lesion, indicating that the iSPN cell ensembles active on the different movement types were less spatially distinguishable after 6-OHDA lesion. $P = 0.02$; Kolmogorov–Smirnov test. Plots in **d**–**g** are based on $n = 102$ and $n = 126$ comparisons between bouts of the same movement type, and 255 and 315 comparisons of bouts of different movement types, in 17 *Drd1a*^{cre} and 21 *Adora2a*^{cre} mice, respectively, before 6-OHDA lesion, and on $n = 72$ and $n = 78$ comparisons between bouts of the same movement type, and 180 and 195 comparisons of bouts of different movement types in 12 *Drd1a*^{cre} and 13 *Adora2a*^{cre} mice, respectively, >14 days after 6-OHDA lesion. The pre-lesion data are the same as in Extended Data Fig. 3. The data >14 days after the lesion are aggregated from the 1-h recordings on day 14 plus the 30-min recordings on day 20 that occurred after saline vehicle injection but before administration of 6 mg kg^{-1} L-DOPA (see Fig. 1a).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



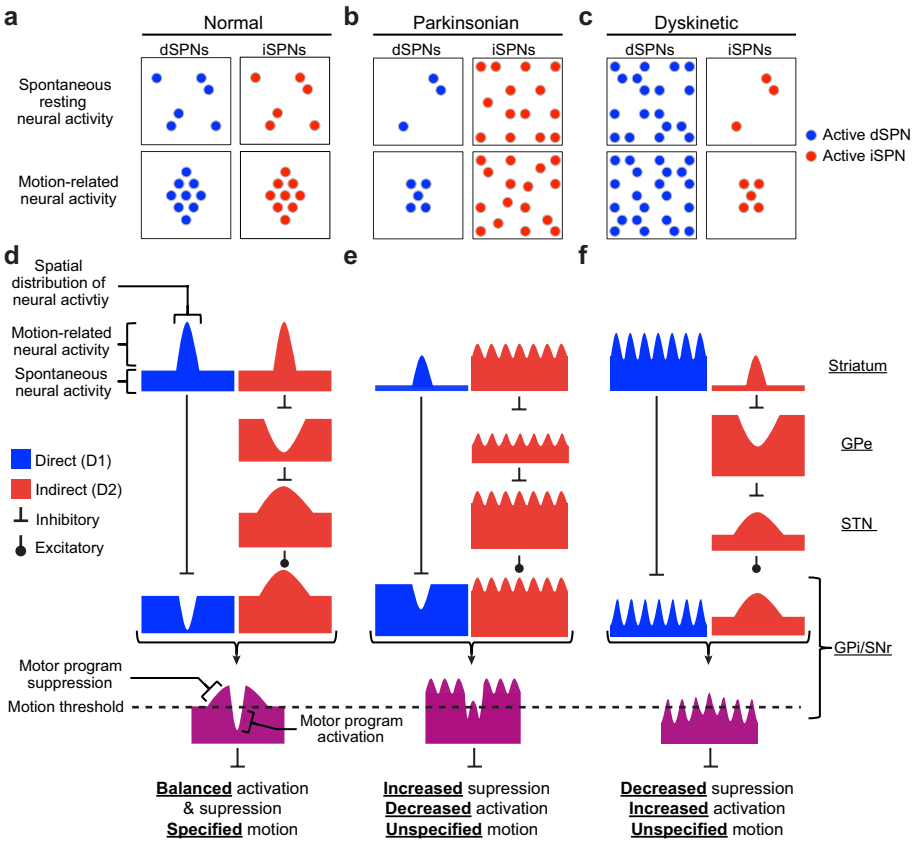
Extended Data Fig. 9 | After unilateral lesion of SNc dopamine cells, L-DOPA and dopamine agonists induce a contralateral turning bias, and L-DOPA can also induce dyskinesia. a–c, Time traces showing mean ± s.e.m. effects of different doses of L-DOPA, SKF81297 and quinpirole on the contralateral rotational bias of mice, scored in 20-min time bins, >14 days after unilateral 6-OHDA lesion. All drug doses had their maximal behavioural effects approximately 30 min after drug administration. All dose–time interactions differed significantly from vehicle treatment ($P < 5 \times 10^{-3}$), and the effects of L-DOPA and SKF81297 were dose-dependent ($P < 5 \times 10^{-3}$). Two-way, repeated-measures ANOVA. **d**, Time traces of the mean ± s.e.m. mouse rotational bias for the three different drugs in **a**–**c**, at doses of each drug that induce comparable levels of rotational bias during the time window used in our studies for Ca^{2+} imaging (shaded in grey), during which we assessed the effects of these drugs on dSPN and iSPN activity in parkinsonian mice (Fig. 4). **e**, Comparison of the net rotational bias in the first 40 min after administration of the three drugs at the same doses as in **d**. Note that, as assessed by their effects on turning behaviour, these three compounds were statistically indistinguishable at these doses. 'ns' denotes $P > 0.2$ for comparisons between drugs; $n = 13$ mice; Wilcoxon signed-rank test. However, all drugs had significant effects relative to vehicle treatment. $^{**}P < 5 \times 10^{-3}$; Wilcoxon signed-rank test; $n = 13$ mice. **f**–**i**, Time traces showing how different dosages of L-DOPA affect the mean ± s.e.m. axial (**f**), orofacial (**g**), limb (**h**) and total (**i**) AIMS values, a behavioural measure of dyskinesia in preclinical models of Parkinson's disease[31]. All three L-DOPA doses had significant effects compared to vehicle treatments ($P < 0.02$; two-way ANOVA) and differed significantly in their effects at the different dosages ($P < 0.05$). As in **d**, the grey shading in **i** denotes the time window used for Ca^{2+} imaging (Figs. 4, 5). **j**, L-DOPA increased the peak total AIMS values, in a dose-dependent manner, in the first 40-min after mice received the L-DOPA doses in **f**–**i**. $^{*}P < 0.05$, $^{**}P < 5 \times 10^{-3}$; Wilcoxon signed-rank test. Error bars in **a**–**d** and **f**–**i** indicate s.e.m. All data in **a**–**j** are from a distinct cohort of $n = 13$ freely moving mice not subject to the protocol of Fig. 1a.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | SPNs encode specific movements via the activation of spatial clusters of cells consisting of both dSPNs and iSPNs. a, Classical models of the basal ganglia, such as the rate model[1,2], posit opposing roles for dSPNs and iSPNs in movement generation. A simple instantiation of these ideas (a 'stop–go' model[11]) predicts increased dSPN and decreased iSPN activity at movement initiation, and decreased dSPN and increased iSPN activity at movement termination. This is depicted here schematically in the traces of aggregate SPN Ca^{2+} activity in the model (middle), relative to motion onset and offset (top). However, in the actual data the activity of both SPN types increases at motion onset and decreases at motion offset (bottom). **b**, Other types of models are more compatible with the concurrent activation of both SPN types at motion onset. In 'suppression–selection' models[15], dSPNs select a motor program, while iSPNs simultaneously suppress competing motor programs. In these models, specific subsets of dSPNs and iSPNs coactivate to select and suppress specific motor programs, respectively. Here we depict this idea using schematic traces of single-cell activity. dSPNs that encode either 'type A' or 'type B' movements activate selectively during one or the other type of movement (see top trace of movements in **a**). By contrast, iSPNs, which suppress competing movements such as those of 'type C' and 'type D', activate on both 'type A' and 'type B' movement bouts. A prediction of this model is that the cell populations encoding different types of movement are more similar (that is, less selective), and possibly more broadly distributed anatomically, for iSPNs than dSPNs. However, the actual data reported here show that the selectivity and spatial distributions of neural responses during movement are comparable in both SPN types (Fig. 2c–e and Extended Data Figs. 3d–f, 4j, k). **c**, More specifically, we found action-selective ensembles in both dSPNs and iSPNs (Extended Data Fig. 3). Unlike the suppression–selection model in **b**, our findings suggest that dSPNs and iSPNs coordinate to select motor programs, depicted here as traces of single-cell activity in which movement 'type A' and 'type B' encoding dSPNs and iSPNs activate selectively on bouts of movement types A and B, respectively (see top trace of movements in **a**). Furthermore, the dSPNs or iSPNs activated for one movement type were physically closer to each other than to SPNs activated for different movement types (Extended Data Fig. 3f). These analyses quantify what is visually apparent in the Ca^{2+} videos of striatal activity, which show that multiple, segregated clusters of coactive dSPNs or iSPNs activate at the onset of motion (Supplementary Videos 1, 2). Finally, using two-photon microscopy to image the simultaneous patterns of Ca^{2+} activity in dSPN and iSPN ensembles, we found that the coactivity of nearby dSPN–iSPN pairs was indistinguishable from that of pairs of the same SPN type (Extended Data Fig. 4). Combining these findings, our data show that different subsets of spatially intermingled dSPNs and iSPNs activate for distinct types of movement (for example, 'type A' and 'type B' movements, bottom depictions of 'spatial distribution').

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 11 | Implications for basal ganglia dysfunction in parkinsonian and dyskinetic states. a–c, Schematic depictions of the spatial organization of SPN activity. Active cells are shown as filled circles. **a**, Prior to dopamine loss both dSPNs and iSPNs exhibit spatially clustered patterns of coactivation when mice are resting. Clusters of the two cell types are spatially overlapping (Extended Data Figs. 4, 10). During movement, both cell types undergo an increase in the number of active cells and extent to which activity is spatially clustered. **b**, After dopamine loss dSPN activity decreases in rate but remains spatially clustered in resting mice (Fig. 3e, j). Moreover, dSPNs still coactivate in a spatially clustered manner during movement initiation (Fig. 3g). The rate of iSPN activity is increased after dopamine depletion in resting mice (Fig. 3e). However, iSPNs are less responsive at the onset of and throughout movement, and their motion-related activity shows almost no spatial clustering (Fig. 3g–j). **c**, During L-DOPA-induced dyskinesia (LID), dSPNs are hyperactive, unresponsive to movement, and spatiotemporally decorrelated (Fig. 5c–h), much like iSPNs after dopamine loss. During LID, iSPNs are hypoactive but retain their movement-related activation and spatial clustering (Fig. 5c–h), much like dSPNs after dopamine depletion. **d–f**, Iconic depictions of the SPN activity patterns during normal, parkinsonian and dyskinetic states, and the implications for downstream basal ganglia function. Icons are inspired by those in previous work[32]. For each icon in the top row, the height of the rectangular pedestal represents the amplitude of spontaneous SPN activity in resting mice. The height and width of protrusions above the pedestals respectively denote the amplitude and spatial distribution of motion-related SPN activity. On the basis of the known excitatory and inhibitory projections between basal ganglia nuclei, the icons in subsequent rows show the predicted consequences for the downstream neural targets. In these icons, indentations into the rectangular pedestal denote inhibitions of neural activity during active movement. The globus pallidus pars externa (GPe) and the sub-thalamic nucleus (STN) are in the basal ganglia's indirect pathway and receive signals from striatal iSPNs. In the direct pathway, dSPNs project directly to the globus pallidus pars interna and substantia nigra pars reticulata (GPi/SNr), where signals from the two basal ganglia pathways converge (purple icons). GPi/SNr activity influences motor program selection by modulating downstream thalamacortical neurons and brainstem motor nuclei (not depicted). **d**, Before dopamine depletion, GPi/SNr cells receive signals during movement from the direct and indirect pathways that should induce, respectively, spatially structured patterns of neural activation and inhibition in GPi/SNr. We hypothesize that the convergence and spatial patterning of these bidirectional modulations enhance the specificity of the disinhibitory signals transmitted from GPi/SNR (left purple icon). In this view, the spatially coordinated, joint modulation of activity in GPi/SNr by both dSPNs and iSPNs may be crucial for well-choreographed motor programs[15]. **e**, After dopamine depletion, spontaneous activity is reduced in dSPNs and increased in iSPNs. The rate model predicts that these changes in spontaneous activity lead to increases in spontaneous GPi/SNr activity (middle purple icon), tonically suppressing movement. Dopamine depletion also abolishes the spatially clustered, motion-related activation of iSPNs. This added deficit in motion-related iSPN activity is unaccounted for by the rate model and may upset the normal bidirectional spatial patterning of GPi/SNR output activity. This in turn may reduce the specificity of GPi/SNr inhibition, causing abnormal movement specification and motor coordination. **f**, During LID, spontaneous activity rises in dSPNs and declines in iSPNs. The rate model predicts that these changes should tonically drive movement by suppressing GPi/SNr activity (right purple icon). LID also disrupts the amplitude and spatial structure of motion-related dSPN activity. This decorrelated dSPN hyperactivity may disrupt the normally focused suppression of GPi/SNr activity during movement, resulting in increased but less specified movements.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0078-2

A midline thalamic circuit determines reactions to visual threat

Lindsey D. Salay[1], Nao Ishiko[1] & Andrew D. Huberman[1,2,3,4]*

How our internal state is merged with our visual perception of an impending threat to drive an adaptive behavioural response is not known. Mice respond to visual threats by either freezing or seeking shelter. Here we show that nuclei of the ventral midline thalamus (vMT), the xiphoid nucleus (Xi) and nucleus reuniens (Re), represent crucial hubs in the network controlling behavioural responses to visual threats. The Xi projects to the basolateral amygdala to promote saliency-reducing responses to threats, such as freezing, whereas the Re projects to the medial prefrontal cortex (Re→mPFC) to promote saliency-enhancing, even confrontational responses to threats, such as tail rattling. Activation of the Re→mPFC pathway also increases autonomic arousal in a manner that is rewarding. The vMT is therefore important for biasing how internal states are translated into opposing categories of behavioural responses to perceived threats. These findings may have implications for understanding disorders of arousal and adaptive decision-making, such as phobias, post-traumatic stress and addictions.

Vision is a rich source of information about the presence of impending threats, such as predators or heights. To ensure survival, visual threats must be perceived and integrated with the neural circuits that control internal states (for example, arousal level) in order to engage the appropriate behavioural responses. However, where in the brain this integration occurs is unknown. Adaptive behavioural responses to visual threats fall into two general categories: those that reduce saliency to predators (to avoid detection), such as freezing or hiding[1], and those that enhance saliency (to avoid capture), such as fleeing or retaliatory aggression[2–4].

Visual threat activates midline thalamic nuclei

Fear, which manifests as a sudden shift towards freezing or hiding, can be evoked in mice by exposing them to a dark, rapidly expanding, overhead 'looming' stimulus that resembles an approaching predator[1] (Fig. 1a, Supplementary Videos 1 and 2). Fear does not occur in the absence of looming, or if looms are presented in the lower visual field[1,5] (Fig. 1b). In response to overhead looming, mice ($n = 15$, controls) spent 91.9% of the loom presentation time either freezing, or hiding in the provided shelter. For the remaining fraction of time (8.1%), mice were either ambulatory (walking; 6.4%) or running (1.7%), the vast

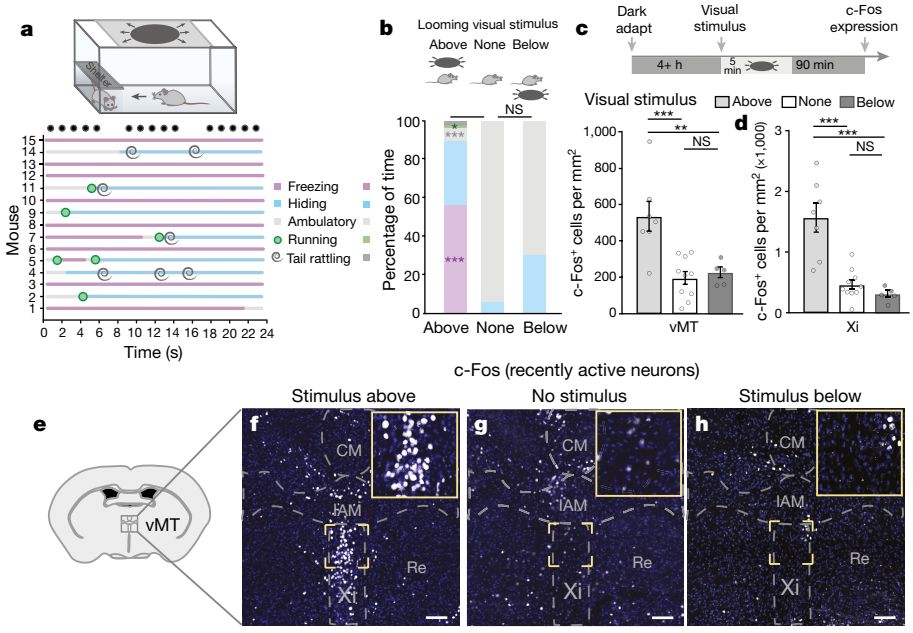


Fig. 1 | Visual threat activates the ventral midline thalamus. a, Ethogram of responses to a looming threat. Black circles represent the looming stimulus (15 expansions in 24 s). **b**, Behaviours in response to looming from above, no loom and looming from below. **c**, **d**, Looming from above, but not below, induces vMT c-Fos activation (**c**) and Xi activation (**d**) relative to no loom. **e**, Location of the ventral midline thalamus within the mouse brain. **f**–**h**, c-Fos$^+$ (white) neurons in the vMT after overhead looming (**f**), no looming stimulus (**g**) or looming from below (**h**). Scale bars, 100 μm. CM, central medial nucleus; IAM, intrantereomedial nucleus; Re, nucleus of reuniens; Xi, xiphoid nucleus. Data are mean ± s.e.m. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$; NS, not significant. See Supplementary Table 1 for further details of the statistical analyses.

[1]Department of Neurobiology, Stanford University School of Medicine, Palo Alto, CA, USA. [2]Stanford Neurosciences Institute, Stanford University School of Medicine, Palo Alto, CA, USA. [3]BioX, Stanford University School of Medicine, Palo Alto, CA, USA. [4]Department of Ophthalmology, Stanford University School of Medicine, Palo Alto, CA, USA. *e-mail: adh1@stanford.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

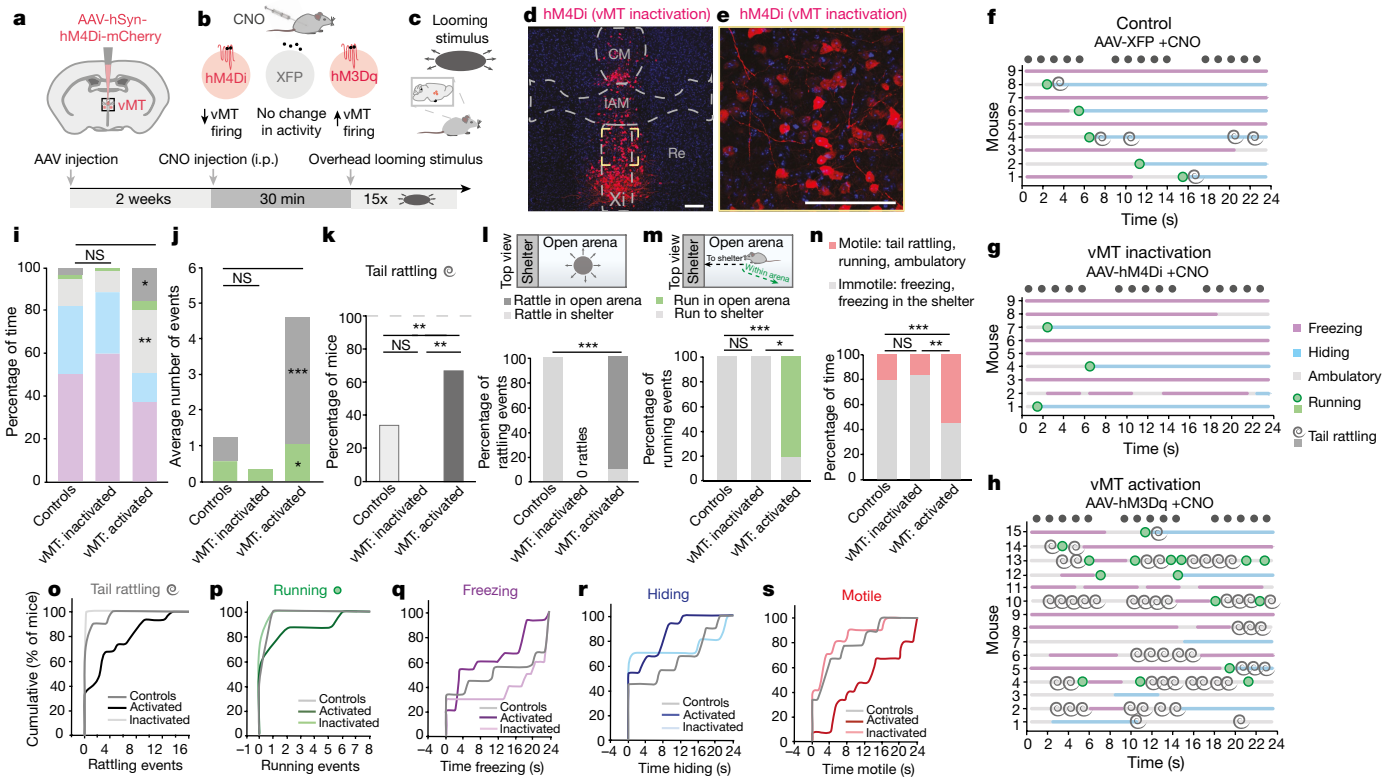


Fig. 2 | vMT neurons modulate visual threat responses. a–c, Diagrams depicting different stages of the experimental protocol. i.p., intraperitoneal. **d**, hM4Di–mCherry^{+} cells (red) in the vMT. **e**, A zoomed-in view of the region outlined by a yellow square in **d**. Scale bars (**d**, **e**), 100 μm. **f–h**, Ethograms of responses to overhead looming in control mice treated with CNO (**f**), mice with vMT inactivation (**g**) and mice with vMT activation (**h**). **i–m**, Effects of vMT activation or inactivation on the behavioural responses of mice. vMT inactivation eliminated tail rattling (**i–l**); vMT activation increased tail rattling (**i**, duration; **j**, incidence; **k**, percentage of mice) and running (**i**, duration; **j**, incidence) in the open arena (**l**, **m**). vMT activation increased ambulatory time in the open (**i**) and total motile behaviours (**n**) relative to control mice. **o–s**, Cumulative frequency distribution plots of behavioural responses. Data are mean ± s.e.m. $*P<0.05$, $**P<0.01$, $***P<0.001$; NS, not significant. See Supplementary Table 1 for further details of the statistical analyses.

majority of which was in the direction of the shelter in order to hide. Control mice spent a very small portion (2%) of the total overhead loom presentation time 'tail rattling', a behaviour typically associated with aggression[6–8]. Notably, all of the tail rattling occurred when the mice were hiding under the shelter, not in the open, unsheltered portion of the arena (Fig. 1a).

To parse the neural circuits that determine behavioural responses to threatening visual stimuli, we examined c-Fos, a marker of recently active neurons, in the brains of mice recently exposed to overhead looming. Neurons in the vMT were highly activated by looms presented overhead, but not by looms presented below (Fig. 1c–h), with the greatest activation occurring in the Xi (Fig. 1d–h). Because the vMT resides outside the primary visual pathway, it may represent an important station for modulating and/or promoting behavioural responses to visual threats.

vMT neurons modulate threat responses

To further assess the functional role of the vMT, mice were stereotaxically injected with an adeno-associated virus (AAV) expressing hM4Di ($G_{i/o}$-coupled human muscarinic M_4 designer receptor exclusively activated by a designer drug (DREADD)) tagged with mCherry (AAV-hSyn-hM4Di-mCherry) ($n=9$ mice; Fig. 2a–e). Control mice ($n=9$) received injections of a virus lacking DREADD (for example, AAV-hSyn-mCherry). Other mice ($n=15$) received injections of AAV-hSyn-hM3Dq-mCherry (G_q-coupled human muscarinic M_3 DREADD) into the vMT. hM4Di and hM3Dq receptors bind clozapine *N*-oxide (CNO) and cause robust decreases or increases, respectively, in the firing of central-nervous-system neurons[9,10] (Fig. 2b). vMT neurons expressing hM4Di and exposed to CNO showed a significant decrease in c-Fos activation, whereas a significant increase in c-Fos activation occurred in vMT neurons expressing hM3Dq (Extended Data Fig. 1).

Marked differences in responses to looming were observed between the three groups of mice. Control mice displayed normal ratios of behaviours, with freezing and hiding accounting for the vast majority of the loom-stimulus presentation time (Fig. 2f, i, o–r, Extended Data Fig. 2). Mice in which hM4Di–mCherry was expressed in the vMT and CNO had been administered (that is, mice in which the vMT was inactivated), showed normal responses to overhead looming stimuli by freezing or running to hide (Fig. 2g, i, p–r); however, the overall number of mice exhibiting tail rattling was reduced from 3 out of 9 control mice to 0 out of 9 vMT-inactivated mice (Fig. 2f, k, Supplementary Video 3).

Considerable shifts in the behavioural responses to overhead looms were observed in mice that expressed hM3Dq and were treated with CNO (that is, mice in which the vMT was activated). Although the mice hid at normal frequency (Fig. 2h, i, r), they exhibited significant increases in the number (Fig. 2h, j, o) and duration (Fig. 2i) of tail-rattling episodes, most of which occurred outside the confines of the shelter (Fig. 2h, l, Supplementary Video 4). The overall percentage of mice exhibiting tail rattling was also significantly increased (Fig. 2k). Tail rattling is associated with aggression in male mice before fighting[7,11] and in female mice eliciting maternal territorial behaviour[12], although it can reflect non-aggressive states as well[11,13]. We observed significant increases in tail rattling in both male and female mice, with no sex differences (Extended Data Fig. 2). Although the emotional state associated with tail rattling is unclear, it unequivocally increases the visual and auditory saliency of the mouse to predators (Supplementary Video 4). Mice treated with both hM3Dq and CNO also displayed

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

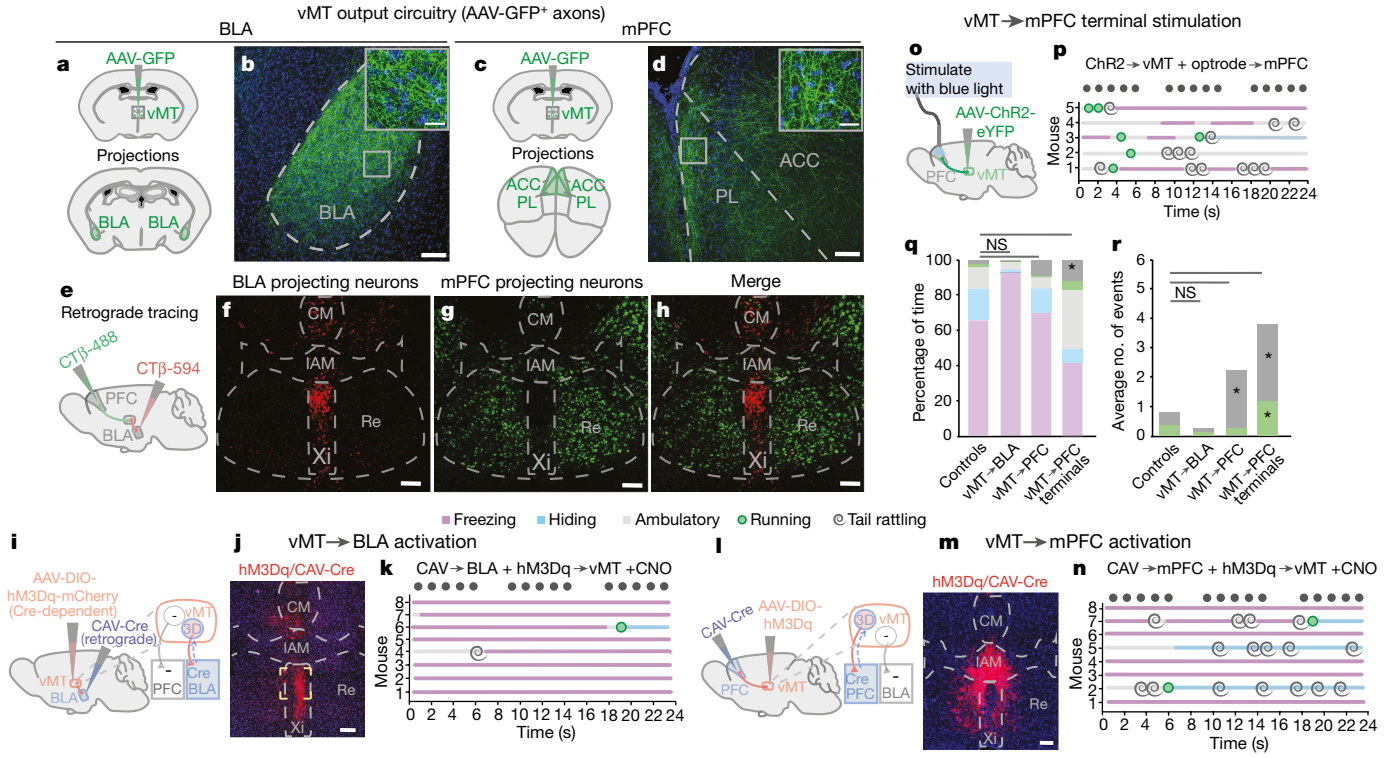


Fig. 3 | Divergent vMT outputs influence threat responses. a–d, Mice were injected with AAV-GFP in the vMT (**a**, **c**; $n = 17$ mice) and axons (green) were observed in the BLA (**b**) and mPFC (**d**). **e**, Locations in the brain at which retrograde tracing was performed. **f–h**, Retrograde labelled vMT→BLA (red, **f**), vMT→mPFC projection neurons (green, **g**), and the merged image (**h**); $n = 5$ mice. **i**, **j**, Injections (**i**) and representative image of hM3Dq/Cre$^+$ neurons (red) in the vMT, projecting to the BLA (**j**). **k**, Ethogram of the responses to looming in mice when the vMT→BLA neurons are activated. **l**, **m**, Injections (**l**) and representative image of hM3D/Cre$^+$ neurons (red) in the vMT that project to the mPFC (**m**). **n**, Ethogram of the responses to looming in mice with vMT→mPFC neurons activated. **o**, **p**, Schematic of injections (**o**) and ethogram of responses to looming in mice with vMT axon terminals activated in the mPFC (**p**). **q**, **r**. Activating the vMT→mPFC pathway, but not the vMT→BLA pathway, increases tail rattling (**q**, duration; **r**, incidence) relative to controls. BLA, basolateral amygdala; mPFC, medial prefrontal cortex; PL, prelimbic cortex; ACC, anterior cingulate cortex. Scale bars, 100 μm. Data are mean ± s.e.m. $*P < 0.05$, $**P < 0.01$; NS, not significant. See Supplementary Table 1 for further details of the statistical analyses.

significantly more walking (duration) and running (incidence), with the vast majority of this locomotion occurring within the open arena (Fig. 2 i, j, m, p). Overall, we observed a significant shift towards more motile behaviours in mice treated with both hM3Dq and CNO compared to control mice (Fig. 2n, s), and a shift away from immotile, saliency-reducing behaviours (for example, freezing and hiding). No increase in the amount of motile behaviour occurred in the absence of visual threats (Extended Data Fig. 6). Therefore, vMT influences action selection in response to threats, but does not influence locomotor behaviour in itself.

Distinct threat responses by vMT outputs

Next we explored the structure and function of vMT output circuitry. We injected AAV-XFP into the vMT ($n = 12$ mice; Fig. 3a, c, Extended Data Fig. 3a, b) and two weeks later, collected and analysed the brains (Fig. 3a–d). The two major outputs of the vMT were the basolateral amygdala (BLA) (Fig. 3a, b) and the medial prefrontal cortex (mPFC) (Fig. 3c, d). Two-colour retrograde tracing (Fig. 3e) from these targets revealed that they originate from non-overlapping, spatially compartmentalized sets of vMT neurons. The vMT→BLA pathway originated from neurons in the Xi nucleus (Fig. 3f, h), whereas the vMT→mPFC pathway originated from the Re nucleus (Fig. 3g, h).

The rodent BLA and mPFC have been shown to be involved in a range of functions including the control of fear[14,15] and anxiety-related behaviours[16–18]. Given the divergent outputs of different vMT neurons to each of these targets (Fig. 3e–h), we investigated how each pathway affects the responses to looming stimuli. To study the vMT→BLA pathway, we injected the BLA with a canine adenovirus-2 expressing Cre-recombinase (CAV-Cre), which infects synaptic terminals and thus 'retrogradely' infects neurons with Cre[19]. We then injected AAV-hSyn-DIO-hM3Dq-mCherry (Fig. 3i) into the vMT of the same mice, thereby inducing expression of hM3Dq–mCherry selectively in Cre-expressing vMT→BLA neurons ($n = 8$ mice; Fig. 3j); we then treated the mice with CNO to activate vMT→BLA circuitry. This caused a marked increase in the duration of freezing in response to looms, and a reduction in the amount of time that the mice spent hiding (Fig. 3k, q). No increases in tail rattling were observed after activation of vMT→BLA neurons (Fig. 3k, q, r).

In other mice ($n = 8$) we injected the mPFC, with CAV-Cre, followed by an injection of AAV-hSyn-DIO-hM3Dq-mCherry into the vMT (Fig. 3l); this resulted in specific infection of the hM3Dq-mCherry/Cre$^+$ neurons that project to mPFC (vMT→mPFC; Fig. 3m, Extended Data Fig. 4). We then injected these mice with CNO to selectively activate vMT→mPFC circuitry. This caused a significant increase in tail-rattling events compared to controls, but did not change freezing or hiding behaviours (Fig. 3n, q, r).

In both the vMT→mPFC-activated and the vMT→mBLA-activated mice, we observed mCherry$^+$ axons in the nucleus accumbens (NAc), suggesting that axonal projections from the vMT bifurcate to innervate the NAc. Anterograde tracing from the vMT with AAV-GFP confirmed this (Extended Data Fig. 3a–d). We tested whether the vMT→NAc pathway altered behavioural responses to looming (Extended Data Fig. 3e, f), but found that it did not (Extended Data Fig. 3g). Therefore, the vMT→mPFC and the vMT→BLA circuits determine the behavioural responses to visually threatening stimuli.

To specifically activate the vMT→mPFC pathway and not the bifurcated NAc axons, we injected AAV-ChR2-eYFP (ChR2, channelrhodopsin-2; eYFP, enhanced yellow fluorescent protein)[20] into the

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

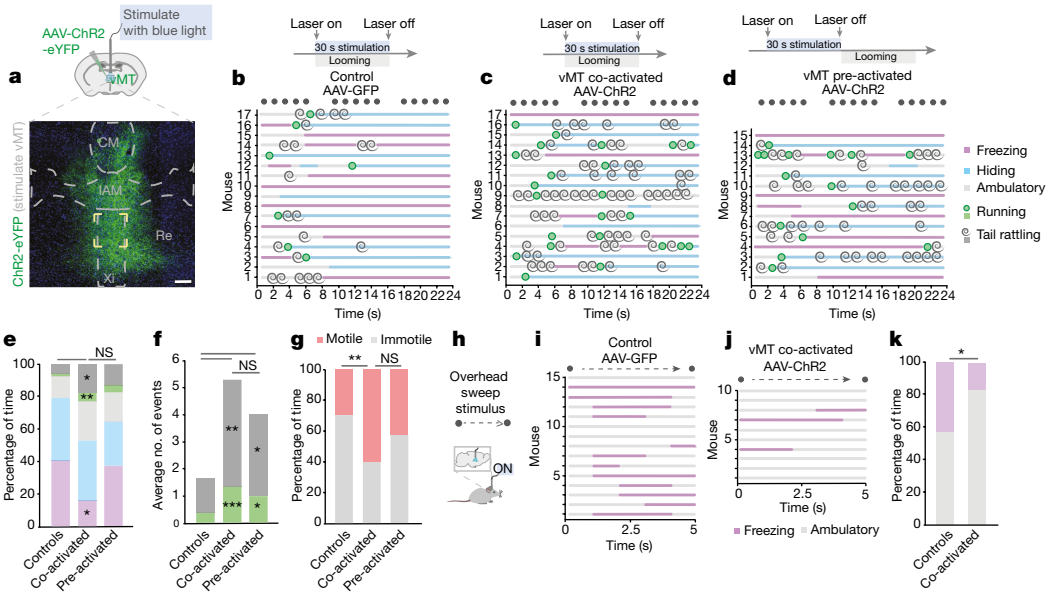


Fig. 4 | Saliency-enhancing behaviours to threat persist after vMT stimulation. a, Representative image of ChR2–eYFP$^+$ cells in the vMT. Scale bars, 100 μm. **b**–**d**, Ethograms of responses to looming in control mice (**b**), mice with vMT activated in concert with the loom (**c**, co-activated) and mice with vMT activated before the loom (**d**, pre-activated). **e**–**g**, Co-activating and pre-activating the vMT increased tail rattling and running (**e**, duration; **f**, incidence). Co-activating the vMT decreased freezing (**e**) and increased total motile behaviours (**g**). **h**, Schematic of the overhead-sweep-stimulus experiment. **i**–**j**, Ethograms of responses to the overhead sweep stimulus in control mice (**i**) and mice in which the vMT is activated during the sweep (**j**). **k**, Co-activating the vMT decreased freezing in response to the sweep. Data are mean ± s.e.m. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. See Supplementary Table 1 for further details of the statistical analyses.

vMT and implanted a fibre-optic cannula above the mPFC (Fig. 3o). Upon optogenetic stimulation of the vMT→mPFC pathway, 100% of mice exhibited tail rattling (Fig. 3p, Extended Data Fig. 5). The stimulation of vMT→mPFC terminals significantly increased tail rattling (duration and incidence) and running (incidence) (Fig. 3p–r). Collectively, these data reveal that each of the two divergent outputs of the vMT determines a unique set of responses to visually threatening stimuli: the vMT→mPFC circuit increases tail rattling and motile behaviours, and the vMT→BLA circuit suppresses these behaviours and encourages freezing instead, effects that do not simply reflect differences in the number of infected neurons in each condition (Extended Data Fig. 4a–d).

Impact of vMT activation prior to loom exposure

To understand further how the vMT affects behavioural response selection to visual threats, we again used virally-expressed ChR2 ($n = 32$ mice; Fig. 4a) and varied the timing of vMT activation relative to loom presentations. In the first condition, the vMT was stimulated at the same time as the presentation of the loom stimulus (Supplementary Video 5). Control mice with AAV-GFP expressed in the vMT, but no ChR2, showed normal loom responses ($n = 17$ mice; Fig. 4b, Extended Data Fig. 2). By contrast, mice with ChR2 expressed and blue light stimulation presented in concert with the looming stimulus exhibited significant increases in both the duration and incidence of tail rattling and running (Fig. 4c, e, f). Overall, vMT stimulation together with looming significantly increased saliency-enhancing behaviours, and decreased saliency-reducing behaviours such as freezing (Fig. 4e, g).

Next we asked whether vMT stimulation before looming also induces shifts in the behavioural responses of mice. We stimulated the vMT for 30 s before any exposure to looming stimuli (Fig. 4d) (n = 15). This induced a remarkably similar pattern of effects to vMT stimulation in concert with looms; in both cases, the mice displayed significantly increased saliency-enhancing behaviours, such as tail rattling and running, compared to controls (Fig. 4e–g). This suggests that vMT stimulation causes a shift in the internal state of the mouse and does not somehow alter visual perception in itself.

To further test the idea that vMT activation shifts mice towards states that favour saliency-enhancing responses to threat, we exposed a subset of mice to a visual 'sweep' stimulus, which reliably induces freezing but not escape and/or hiding behaviours[21] (Fig. 4h, i, k). Activating the vMT with ChR2 and blue light during the presentation of visual sweeps significantly reduced the amount of sweep-induced freezing and promoted saliency-enhancing behaviours (Fig. 4j, k). ChR2-induced activation of the vMT also significantly reduced freezing and avoidance of olfactory threats[22,23], and increased saliency-enhancing ambulatory and approach behaviours (Extended Data Fig. 6a–c).

vMT activation increases arousal level

The fact that vMT activation causes robust shifts in the responses to fear-inducing stimuli even when it is induced before the stimuli are delivered suggests that vMT activation modulates arousal state. To explore this, we examined the size of the pupil in response to illumination at constant light levels[24], which is a fast read-out of arousal[25,26].

Heightened arousal increases pupil diameter[26]. We explored how vMT stimulation affects pupil size at constant 100-lx ambient light. vMT stimulation resulted in a significant increase in pupil size (60% change) compared with that of control mice (Fig. 5a–c). Even after vMT stimulation was ceased, the pupil remained significantly dilated (80%; Fig. 5c). To separate the arousal-driven from the light-driven component of the pupil reflex, we tested the consensual[27] reflex of the pupil in one eye, in response to light pulses delivered to the opposite eye in otherwise dark-adapted mice. Mice in which vMT activity was enhanced showed normal pupil constriction in response to light, but in the dark showed significantly increased pupil dilation compared to controls (Fig. 5d, e). Increased dilation was also observed in mice in which the vMT→mPFC pathway was stimulated, but not in mice in which the vMT→BLA pathway was stimulated (Fig. 5f). Sham stimulation and CNO alone did not affect pupil size (Extended Data Fig. 7). vMT activation also caused a significant increase in heart rate (Fig. 5g, h) and a trend towards increased breathing rate (Fig. 5i). Thus, pan-vMT and vMT→mPFC activation does not affect light-driven pupil responses; rather, it increases arousal.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

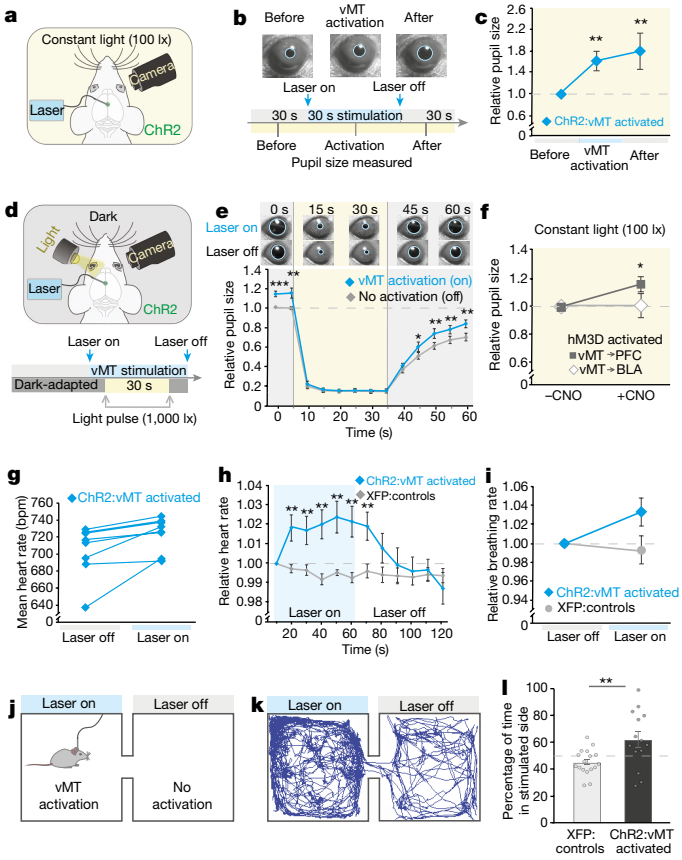


Fig. 5 | vMT activation increases arousal. a, Schematic of experimental set-up to analyse arousal-driven pupillary responses to vMT stimulation. **b**, **c**, Pupil size in constant light (100 lx) conditions before, during and after vMT stimulation ($n = 12$ mice) relative to controls ($n = 8$ mice, dashed line in **c**). **d**, Schematic of experimental set-up to analyse light-driven pupillary responses in dark-adapted mice. **e**, Relative pupil size before, during and after a 30-s light pulse (1000 lx) in mice with and without vMT stimulation ($n = 11$ mice). **f**, Under constant light conditions, activating the vMT→mPFC pathway, but not the vMT→BLA pathway increased relative pupil size. **g**–**i**, Heart rate (**g**, **h**) and breathing rate (**i**) in mice with and without vMT stimulation ($n = 8$ mice). **j**, Illustration of the RTPP test. **k**, Representative tracing of a mouse with vMT stimulation in the RTPP test. **l**, Mice with vMT activation ($n = 14$) spent more time in the stimulated side as compared to control mice ($n = 17$). Data are mean ± s.e.m. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. See Supplementary Table 1 for further details of the statistical analyses.

Changes in arousal can be associated with positive or negative states. We questioned whether vMT stimulation is rewarding, aversive or neutral. ChR2-expressing mice ($n = 14$) were placed in a two-chamber arena and the vMT was stimulated using a fibre-optic cannula every time the mouse entered one particular side of the chamber (Fig. 5j). The stimulation was ceased when the mice explored the opposite side. This real-time place preference (RTPP) test revealed a strong preference for the vMT-stimulated side (Fig. 5k, l). This did not reflect an alteration in locomotor activity (Extended Data Fig. 6). vMT activation also did not disrupt the perception of danger cues in a visual cliff test, aggression or tail rattling in response to a male intruder, or overall levels of anxiety in an open-field test (Extended Data Fig. 6). Therefore, increases in arousal induced by vMT stimulation are rewarding; that is, they have positive valence.

vMT firing scales with perceived threat

Next we carried out single-unit tetrode recordings from freely behaving mice before, during and after exposure to looming stimuli (Fig. 6a; see Methods). Example spiking activity from five different vMT cells in response to the first round of looming stimuli is shown in Fig. 6b. Four out of five cells showed increased firing to looms relative to the pre-loom condition. Overall, the firing rates of vMT neurons increased by approximately 30% in response to looms on the first day of loom exposure (Fig. 6c), which is when the most robust behavioural responses to loom occur.

Aligning vMT-spiking activity with the behaviours exhibited during looming showed that vMT spiking scaled with the degree to which the mice exhibited different behavioural responses. The firing rates of vMT neurons markedly increased when the mouse tail-rattled, walked or ran, but not when the mouse froze (Fig. 6d). Overall, there was a significant increase in vMT firing when the mouse exhibited saliency-enhancing behaviours as opposed to saliency-reducing behaviours to looming (Fig. 6d–f).

It is well established that mice habituate their responses when looming is repeatedly presented over several days[1,5]. We explored whether vMT firing adjusts to those changes, and found that the relative pre-loom versus loom firing rate of vMT neurons attenuated in parallel with the behavioural attenuation (Fig. 6g), an effect also seen with c-Fos (Extended Data Fig. 8). These data indicate that the firing of vMT neurons scales according to the novelty of a threat stimulus, and in a way predicts the changes in behavioural responses to visual threats over time.

Discussion

We show that when vMT neurons are activated before, or during, simulated predator looms or olfactory threat, mice shift from saliency-reducing to saliency-enhancing reactions to those threats. Whereas looms typically induce freezing, activation of the vMT or the vMT→mPFC pathway induced tail rattling—a behaviour associated with aggression[11–13]. Although the range of emotional states represented by behaviours such as tail rattling is likely to be broad, saliency-enhancing behaviours in general can be thought of as 'courageous' because they increase the visibility of the mouse in the direct presence of a perceived threat. Indeed, from the perspective of predator–prey interactions, avoiding predation reflects a trade-off between avoiding detection (for example, freezing, hiding or camouflage) and avoiding capture (for example, running or fighting); which category of behaviour is adaptive requires ongoing risk assessment. Our data indicate that arousal level appears to bias which category of behavioural response mice will exhibit in response to threats (Extended Data Fig. 9). Activation of the vMT consistently increased arousal and saliency-enhancing 'courageous' behaviours to threats. This is unlikely to indicate increased stress because our real-time place preference test showed that the internal state caused by vMT activation is rewarding (that is, it has positive valence). In light of this, it is notable that at least two previous studies found that self-stimulation of the human brain area that is homologous[28] to the rodent vMT is strongly reinforcing, even more so than activation of areas associated with sexual arousal[29,30].

The vMT structures described here reside well outside the primary visual pathway and receive diverse sources of input from sensory structures, including the superior colliculus[31], as well as from hindbrain areas associated with arousal and emotion, such as the raphe and the periaqueductal grey[32,33] (Extended Data Fig. 10). Thus, the vMT may act as a hub for integrating several types of sensory inputs with internal states in order to select specific categories of behavioural outputs. In contrast to the vast inputs to the vMT, the outputs of the vMT are fairly restricted: they include the basolateral amygdala, the medial prefrontal cortex (Fig. 3) and the nucleus accumbens (Extended Data Fig. 3), structures all shown previously to be associated with fear and/or reinforcement[14,16,34].

Our data point to the powerful influence that activation of thalamic midline structures can have on arousal and behavioural responses to threatening stimuli. Other midline-residing structures—such as the ventromedial hypothalamus and the periaqueductal grey—have been shown by others to control aggression and arousal[35,36], which raises

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

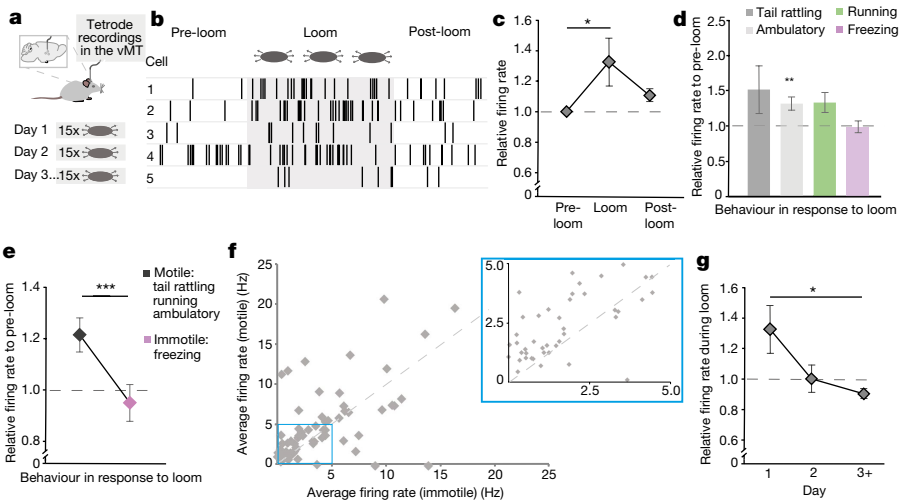


Fig. 6 | vMT neurons are activated during saliency-enhancing behaviours. a, Tetrode recordings in the vMT were performed during presentation of the looming stimulus. **b**, Sample raster plots of five isolated units before, during and after the looming stimulus. **c**–**e**, Mean relative firing rates on the first day of loom presentation (**c**, 23 cells from 4 mice), during different behaviour epochs (**d**, 87 total cells from 4 mice) and during motile and immotile behaviours (**e**, 67 cells from 4 mice), relative to pre-loom (dashed line). **f**, Mean firing rates of all the recorded units during motile behaviour as compared to immotile behaviour in response to looms. Most of the points are above the $y = x$ line. **g**, Mean relative firing rate across several days of looming presentation (87 total cells across 4 mice). Data are mean ± s.e.m. $^{*}P < 0.05$, $^{***}P < 0.001$. See Supplementary Table 1 for further details of the statistical analyses.

ideas about whether midline nuclei of the hind-, mid- and forebrain may represent a larger circuit devoted to sensory-autonomic modulation. In the future, experiments that adjust the activity of several brain areas such as the periaqueductal grey, the ventromedial hypothalamus and the vMT could shed light on this idea. The strong convergence of circuitries involved in arousal, reward and action selection to perceived threats in structures such as the vMT also raises the possibility that dysregulation of midline thalamic function may underlie disorders of these processes such as addiction, phobias and post-traumatic stress disorder.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0078-2.

Received: 19 June 2017; Accepted: 23 March 2018;
Published online 2 May 2018.

1. Yilmaz, M. & Meister, M. Rapid innate defensive responses of mice to looming visual stimuli. *Curr. Biol.* **23**, 2011–2015 (2013).
2. Fanselow, M. & Lester, L. in *Evolution and Learning* (eds Bolles R. C. & Beecher M. D.) 185–212 (Lawrence Erlbaum Associates, Hillsdale, 1988).
3. Eilam, D. Die hard: a blend of freezing and fleeing as a dynamic defense—implications for the control of defensive behavior. *Neurosci. Biobehav. Rev.* **29**, 1181–1191 (2005).
4. Crawford, M. & Masterson, F. A. Species-specific defense reactions and avoidance learning. An evaluative review. *Pavlov. J. Biol. Sci.* **17**, 204–214 (1982).
5. Wei, P. et al. Processing of visually evoked innate fear by a non-canonical thalamic pathway. *Nat. Commun.* **6**, 6756 (2015).
6. Scott, J. P. Agonistic behavior of mice and rats: a review. *Am. Zool.* **6**, 683–701 (1966).
7. Miczek, K. A., Maxson, S. C., Fish, E. W. & Faccidomo, S. Aggressive behavioral phenotypes in mice. *Behav. Brain Res.* **125**, 167–181 (2001).
8. Unger, E. K. et al. Medial amygdalar aromatase neurons regulate aggression in both sexes. *Cell Rep.* **10**, 453–462 (2015).
9. Lim, J.-H. A. et al. Neural activity promotes long-distance, target-specific regeneration of adult retinal axons. *Nat. Neurosci.* **19**, 1073–1084 (2016).
10. Urban, D. J. & Roth, B. L. DREADDs (designer receptors exclusively activated by designer drugs): chemogenetic tools with therapeutic utility. *Annu. Rev. Pharmacol. Toxicol.* **55**, 399–417 (2015).
11. St John, R. D. Genetic analysis of tail rattling in the mouse. *Nature* **241**, 549–551 (1973).
12. Curley, J. P., Davidson, S., Bateson, P. & Champagne, F. A. Social enrichment during postnatal development induces transgenerational effects on emotional and reproductive behavior in mice. *Front. Behav. Neurosci.* **3**, 25 (2009).
13. Beilharz, R. G. & Beilharz, V. C. Observations on fighting behaviour of male mice (*Mus musculus* L.). *Z. Tierpsychol.* **39**, 126–140 (1975).
14. Corcoran, K. A. & Quirk, G. J. Activity in prelimbic cortex is necessary for the expression of learned, but not innate, fears. *J. Neurosci.* **27**, 840–844 (2007).
15. Adhikari, A. et al. Basomedial amygdala mediates top-down control of anxiety and fear. *Nature* **527**, 179–185 (2015).
16. Martinez, R. C., Carvalho-Netto, E. F., Ribeiro-Barbosa, E. R., Baldo, M. V. & Canteras, N. S. Amygdalar roles during exposure to a live predator and to a predator-associated context. *Neuroscience* **172**, 314–328 (2011).
17. Tye, K. M. et al. Amygdala circuitry mediating reversible and bidirectional control of anxiety. *Nature* **471**, 358–362 (2011).
18. Weible, A. P., Piscopo, D. M., Rothbart, M. K., Posner, M. I. & Niell, C. M. Rhythmic brain stimulation reduces anxiety-related behavior in a mouse model based on meditation training. *Proc. Natl Acad. Sci. USA* **114**, 2532–2537 (2017).
19. Junyent, F. & Kremer, E. J. CAV-2—why a canine virus is a neurobiologist's best friend. *Curr. Opin. Pharmacol.* **24**, 86–93 (2015).
20. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8**, 1263–1268 (2005).
21. De Franceschi, G., Vivattanasarn, T., Saleem, A. B. & Solomon, S. G. Vision guides selection of freeze or flight defense strategies in mice. *Curr. Biol.* **26**, 2150–2154 (2016).
22. Papes, F., Logan, D. W. & Stowers, L. The vomeronasal organ mediates interspecies defensive behaviors through detection of protein pheromone homologs. *Cell* **141**, 692–703 (2010).
23. Root, C. M., Denny, C. A., Hen, R. & Axel, R. The participation of cortical amygdala in innate, odour-driven behaviour. *Nature* **515**, 269–273 (2014).
24. Lucas, R. J., Douglas, R. H. & Foster, R. G. Characterization of an ocular photopigment capable of driving pupillary constriction in mice. *Nat. Neurosci.* **4**, 621–626 (2001).
25. McGinley, M. J., David, S. V. & McCormick, D. A. Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron* **87**, 179–192 (2015).
26. Vinck, M., Batista-Brito, R., Knoblich, U. & Cardin, J. A. Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. *Neuron* **86**, 740–754 (2015).
27. Young, M. J. & Lund, R. D. The anatomical substrates subserving the pupillary light reflex in rats: origin of the consensual pupillary response. *Neuroscience* **62**, 481–496 (1994).
28. Morel, A., Magnin, M. & Jeanmonod, D. Multiarchitectonic and stereotactic atlas of the human thalamus. *J. Comp. Neurol.* **387**, 588–630 (1997).
29. Heath, R. G. Electrical self-stimulation of the brain in man. *Am. J. Psychiatry* **120**, 571–577 (1963).
30. Bishop, M. P., Elder, S. T. & Heath, R. G. Intracranial self-stimulation in man. *Science* **140**, 394–396 (1963).
31. Krout, K. E., Loewy, A. D., Westby, G. W. & Redgrave, P. Superior colliculus projections to midline and intralaminar thalamic nuclei of the rat. *J. Comp. Neurol.* **431**, 198–216 (2001).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

32. Vertes, R. P. & Martin, G. F. Autoradiographic analysis of ascending projections from the pontine and mesencephalic reticular formation and the median raphe nucleus in the rat. *J. Comp. Neurol.* **275**, 511–541 (1988).
33. Van der Werf, Y. D., Witter, M. P. & Groenewegen, H. J. The intralaminar and midline nuclei of the thalamus. Anatomical and functional evidence for participation in processes of arousal and awareness. *Brain Res. Rev.* **39**, 107–140 (2002).
34. Stuber, G. D. et al. Excitatory transmission from the amygdala to nucleus accumbens facilitates reward seeking. *Nature* **475**, 377–380 (2011).
35. Gross, C. T. & Canteras, N. S. The many paths to fear. *Nat. Rev. Neurosci.* **13**, 651–658 (2012).
36 Lin, D. et al. Functional identification of an aggression locus in the mouse hypothalamus. *Nature* **470**, 221-226 (2011).

Acknowledgements This work was supported by the National Institutes of Health (NIH) National Eye Institute grant R01 EY022157 (A.D.H.), by NIH U01 NS090562 (A.D.H.) and by a National Science Foundation Graduate Research Fellowship (L.D.S.). We thank L. Giacomo and C. Mallory for advice and assistance with electrophysiology.

Reviewer information *Nature* thanks D. Lin, H. Shin and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions L.D.S. and A.D.H. conceived the design of the study. L.D.S. and N.I. performed the experiments and analysed the data. L.D.S. and A.D.H. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0078-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0078-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.D.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Mice. All experimental protocols were conducted in accordance with the National Institutes of Health (NIH) guidelines for animal research and were approved by the Stanford University and University of California, San Diego Animal Care and Use Committee (IACUC). Mice were housed under a 12-h light:dark cycle with food and water ad libitum. Male and female C57BL/6 mice were 8–12 weeks old at the time of viral injection. All subjects were randomly assigned to different experimental conditions used in this study. Group sample sizes were chosen on the basis of previous studies.

Viruses. AAV-hSyn-hM3D(Gq)-mCherry, AAV-hSyn-hM4D(Gi)-mCherry, AAV-hSyn-DIO-hM3D(Gq)-mCherry, AAV-hSyn-DIO-mCherry, and AAV-hSyn-eGFP vectors were packaged by the Vector Core at the University of North Carolina at Chapel Hill (UNC Vector Core). AAV-Ef1α-DIO-hChR2(H134R)-eYFP was provided by B. K. Lim (UCSD) and K. Deisseroth (Stanford) and was combined with AAV-iCre (Vector BioLabs) to induce opsin expression. ΔG-rabies-mCherry and ΔG-rabies-GFP were produced by the Salk Institute Gene Transfer, Targeting and Therapeutics Core (GT3). CAV2-Cre virus was obtained from The Institute of Molecular Genetics of Montpellier vector core (France). All viral vectors were stored in aliquots at −80 °C until use.

Stereotaxic surgery and drive implantation. Mice were anaesthetized with 1.5–3.0% isoflurane and given the analgesic buprenorphine (subcutaneously, 0.3 mg kg^{-1}). A midline scalp incision was performed to expose the skull. After performing a craniotomy, 0.5 μl of viral vectors (titer, 10^{12} particles per ml) were injected into the vMT (bregma: −1.0 mm, midline: 0.0 mm, dorsal surface: −4.0 mm) using a Nanoject II (Drummond) injector. The virus was slowly infused, and 10 min after infusion the glass pipette (with a resistance of approximately 0.5 MΩ) was retracted. Half of the mice in each cage were randomly assigned to either control or treatment groups. All control mice were treated with the same experimental procedures, but AAV-XFP was injected instead.

To determine the inputs to the vMT, 0.1–0.4 μl of ΔG-rabies-mCherry was injected into the vMT. Mice were then housed in a biosafety room for 4–6 days to allow the rabies virus to infect and express mCherry in presynaptic cells. To determine the outputs, AAV-GFP was injected into the vMT. The entire brain was sectioned and imaged on a Zeiss slide scanner.

For activating vMT output neurons, CAV-Cre was injected bilaterally into the BLA (bregma: −1.3 mm, midline: 3.0 mm, dorsal surface: 4.75 mm), the mPFC (bregma: +2.5 mm; midline: 0.25 mm; dorsal surface 1.5 mm) or the NAc (bregma: +1.3 mm; midline: 1.0 mm; dorsal surface 4.5 mm). In the same mice, AAV-DIO-hM3D(Gq) was injected into the vMT. Mice were tested after at least three weeks of recovery.

To determine whether the vMT projection neurons to BLA and mPFC represent distinct populations, CTβ−488 (green) was injected into the mPFC and CTβ−594 (red) was injected into the BLA of the same mouse. The vMT was examined for cells with both CTβ−488 and CTβ−594. To further determine whether vMT projection neurons bifurcate to the mPFC and the BLA, mice injected with CAV-Cre to either target and a Cre-dependent AAV to the vMT were examined for axons expressing mCherry in the mPFC and the BLA.

For optogenetic experiments, AAV-ChR2 or AAV-GFP was injected into the vMT. Then, a fibre optic cannula (200 μm diameter, 0.22 NA; Doric Lenses) was implanted above the vMT or the mPFC and was affixed to the skull with dental cement. Mice were given four weeks to recover for stimulating in the vMT and six weeks to recover for stimulating in the mPFC.

For tetrode recording from the vMT, mice were implanted with a microdrive with two or four tetrodes (bregma: −1.0 mm, midline: 0.0 mm, dorsal surface: −3.5 mm). Tetrodes composed from 17-μm polyimide-coated platinum–iridium 90–10% wire (California Fine Wire Company) were attached to the microdrive. The electrode tips were cut flat and plated with platinum to reduce impedances to 100–300 kΩ at 1 kHz. The microdrive was secured to the skull with dental cement and jeweller's screws with one screw connected to the microdrive ground. Mice were given three days to recover from surgery before habituation to the headstage and multi-wire tether.

Recording procedures and spike sorting. After surgery, the tetrodes were turned into the vMT in small increments to reach the appropriate depth at the end of each day. Mice were placed in the looming chamber to record pre-, during- and post-loom activity. Because the headstage partially blocked visual perception directly overhead, looming stimuli were presented at an angle. To ensure the same cell was not counted twice, unit activity was collected from sessions with at least 45 μm separation from previous loom sessions. The microdrive was connected to the recording equipment through an AC coupled unity-gain operation amplifier. Spike amplitudes above 40 μV (around two times the level of noise) were collected.

Spikes were sorted manually with offline cluster-cutting software (TINT software, Axona) based on clustering and waveforms. Autocorrelation and cross-correlation functions were used to separate cells. Only cells with a peak firing rate greater than 1 Hz pre-, during- or post-looms were analysed.

Light delivery. Mice with fibre optic implants were connected to an optical fibre (200 μm diameter, 0.22 NA; Doric Lenses) that delivers blue light from a 473 nm laser. Blue light was delivered in 5-ms pulses at 20 Hz for 30 s before or during the entire duration of the looming stimuli. The light power at the fibre tip was about 10–20 mW, ensuring a light intensity of approximately 1 mW mm^{-2} over the entire vMT.

Clozapine *N*-oxide delivery. For chemicogenetics, mice were administered clozapine *N*-oxide (CNO, Toric Bioscience; 1.5 mg kg^{-1} for mice with hM3Dq, 5.0 mg kg^{-1} for mice with hM4Di) by an intraperitoneal injection 30 min before behavioural analyses.

Behavioural analyses. Mice were handled every day for three days before behaviour testing. No significant differences in behaviour responses were observed between males or females tested during night and day. All behaviour was performed during consistent hours in the afternoon or evening. When mice were tested in several behaviour assays, they were given at least 48 h rest between tests. All behaviours were performed and scored blind to treatment group.

Looming behaviour. Mice were placed in a 50 cm × 25 cm × 40 cm glass chamber with a 24-inch LCD monitor stationed on top facing downwards to display the stimulus. The floor and three walls of the chamber were covered with a matte coating (Krylon) to prevent reflections of the looming stimulus. At one end of the box, there was a 12-cm-high black shelving board to provide a shelter under which mice could hide. A top video camera (Yi Action Camera) and a side video camera (Panasonic HC-W850 camcorder in nightshot mode, 30 fps) recorded the mouse's movements during the session. Mice were given 10 min to habituate to the arena. Then, the looming stimulus (black disc on a white background, 15 expansions over 24 s) was initiated when the mouse was in the centre of the arena. All mice were tested only once owing to the fact that they can habituate to the looming stimulus even after one trial of looming presentation. The locomotor behaviours of the mice were recorded and analysed automatically with a video tracking system (Biobserve software).

The behaviour of the mice was characterized on the basis of the following criteria. Freezing was defined as episodes of three or more seconds in which the mouse was completely immobile except for movements from respiration. Running was defined as events in which the speed of the mouse was greater than two times the average speed before stimulus onset. Hiding was defined as episodes in which the mouse's body was completely under the shelter provided. Tail rattling was defined as events in which the mouse rapidly moved its tail back and forth (1 bout of rattling = 1 s of rattling). Ambulation was defined as all other motile behaviours performed in the open arena.

Pupillary light reflex. To test the arousal-driven aspect of the pupillary reflex, mice were adapted to constant, ambient room light (100 lx) for 1 h. All mice were kept unanaesthetized and manually restrained during the experiment. The pupil was recorded under constant light conditions for 30 s before stimulation (baseline), 30 s during stimulation, and 30 s after stimulation. Control mice were treated exactly the same except for 'sham' stimulation.

To test light-driven pupillary reflexes, mice were dark-adapted in their home cages for 1 h before the experiment. A single blue LED (470 nm, 1000 lx) was placed in front on the left eye. The right eye (consensual pupil reflex) was recorded before (baseline), during (constriction) and after (dilation) the 30 s blue light stimulus. Mice were again dark-adapted for 1 h after the first trial. Then, the same mice were tested for pupillary reflex with optogenetic stimulation, chemicogenetic stimulation or sham stimulation. The pupil size was later measured in ImageJ by an experimenter blind to treatment.

Real-time place preference. Mice were placed into a two-chamber acrylic box (40 cm × 20 cm × 25 cm, each chamber 20 cm × 20 cm). Each chamber has different contextual cues on the wall. Mice were monitored for 20 min. When all four paws were in the stimulation chamber, the stimulation was turned on and remained on until the mouse left the chamber. The mice were tracked with automated tracking software (Biobserve) and the time spent in each chamber was recorded.

Home cage vMT activation. Mice were monitored in their home cages before and during vMT activation by chemicogenetic or optogenetic methods. For optogenetic stimulation, a range of laser intensities (2 mW, 5 mW, 10 mW and 20 mW) and a range of stimulation frequencies (5 Hz, 10 Hz and 20 Hz) were tested to determine an optimal stimulation protocol and monitor behaviour. Behaviour was scored and locomotion behaviour was analysed using automated tracking software (Biobserve). No changes in typical behaviours were observed during stimulation in the home cage (data not shown).

Predator odour avoidance. Mice were individually housed and dark-adapted before testing. All behaviour testing was performed under dim red light at night. Medical gauze with trimethylthiazoline (50% TMT, 5 μl), or no odour (clean gauze) was placed in the cage. Behaviour was monitored for 10 min during odour exposure.

Heart rate and breathing rate measurements. Heart rate and breathing rate were recorded at 1 Hz in freely moving mice using a pulse oximeter (MouseOx Plus;

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Starr Life Sciences). Mice were shaved around the neck to enable the red light to pass through to the detector. Recordings were obtained for 10 min in the home cage with intermittent 1-min periods of blue light stimulation. Data were collected using the MouseOx Plus Conscious Applications Software.

Resident–intruder test. Behaviour assays were performed under dim red light in the dark cycle. Male mice were housed with a female for >7 days. The female was removed 1 h before testing. Solitary mice were tested twice for aggression with WT 129/Sv socially housed stimulus mice of similar body weight for 10 min. Aggressive behaviours (attack, tail-rattle) and non-aggressive behaviours (sniffing, grooming intruder) were scored.

Open-field test. Mice were placed in the outer zone of the open-field chamber (50 cm × 50 cm) for 18 min. The 473 nm laser was intermittently turned on and off with 3-min periods. Total time and distance in the centre (36 cm × 36 cm) was analysed using automated tracking software (Biobserve).

Visual cliff. The visual cliff behaviour was analysed in an open-top Plexiglass chamber. Half of the box was protruded from the counter to provide a 3-foot depth. The same checkerboard pattern was displayed on the countertop and the floor below. A 5-cm high platform was stationed in the middle of the box equally intersecting both the shallow side and the deep side. The mouse was placed on the platform and was allowed to choose between the two sides to step down on. If the mouse stepped down on the shallow side with all four paws, the behaviour choice was designated as shallow. All other behaviour choices (0–3 paws on the shallow side) were designated as deep. Each mouse performed the task five times per day to give a total score of percentage choosing the shallow side.

Sweep behaviour. Mice were placed in the same chamber as that used for the looming behaviour with a 24-inch LCD monitor stationed on top facing downward to display the sweep stimulus. Mice were given 10 min to habituate to the arena. Then, the sweep stimulus (2.5-cm diameter black disc on a white background, sweeps across the screen in 4 s) was initiated when the mouse was in the centre of the arena. All mice were tested only once owing to the fact that they can habituate to the sweep stimulus even after one trial. The locomotor behaviours of the mice were recorded and analysed automatically with a video tracking system (Biobserve software).

Immunohistochemistry. After transcardinal perfusion with saline and then 4% paraformaldehyde, brains were collected and postfixed for 24 h. The brains were then transferred to 30% sucrose for cryoprotection and sectioned coronally at 45 μm. If needed, brains were kept at 4 °C overnight with the following antibodies: rabbit-anti-c-Fos to label active neurons or rabbit-anti-DsRed to enhance the mCherry signal. For secondary detection, Alexa Fluor 488 donkey anti-rabbit or Alexa Fluor 594 donkey anti-rabbit were used. Immunostained tissues were imaged with a Zeiss LSM 880 Airyscan confocal microscope or a Zeiss AxioScan microscope (Zeiss, Germany).

c-Fos induction. After behaviour assays, the brains were collected for a standard c-Fos induction protocol. For visual threat experiments, mice were exposed to a visual stimulus for 5 min: looming above, below or no looming stimulus.

To analyse the efficiency of viral infection and cell activation or inactivation, a similar c-Fos induction protocol was used. For optogenetic experiments, a train of 5-ms light pulses at 20 Hz over 3 min was delivered. For chemicogenetics, mice were given CNO to induce the appropriate change in neural activity.

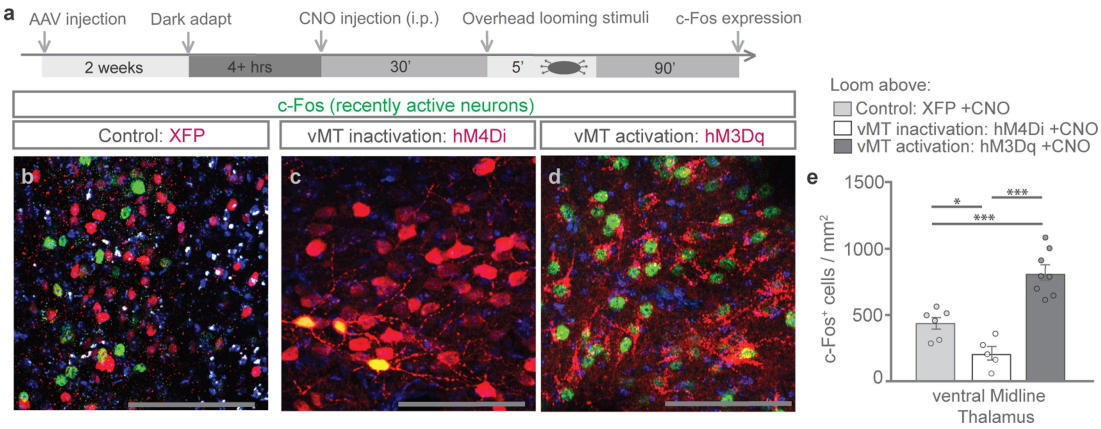
Quantification of cell number. For mice with CAV-Cre injected to the BLA or mPFC and AAV-DIO-hM3D-mCherry into the vMT, we compared the number of $mCherry^+$ cells in the vMT. For c-Fos screens, we compared the number of $c\text{-}Fos^+$ cells in the vMT and Xi in control mice and mice that experienced visual threats. After we immunostained the brains with identical protocols, the brains were imaged on an automated slide scanner (Zeiss). A brain map was overlain on the digital image to identify the appropriate regions using landmark structures, white matter tracts, ventricles and optic tract as reference guides. An experimenter blind to condition analysed the images using the automated ImageJ cell counter.

Statistics and reproducibility. All statistical analyses were performed in R v.3.3.2. (R Development Core Team 2017, https://www.r-project.org/) or GraphPad Prism. The sample sizes were chosen based on common practice in animal behaviour experiments. Data were first tested for normality. If the null hypothesis that the data come from a normal distribution cannot be rejected, parametric tests were used (for example, Student's *t*-test, one-way ANOVA with Tukey's post-hoc test). Otherwise, non-parametric tests were used. Paired tests were used to compare within-group repeated-measures data. Owing to the strong non-normality of the looming behaviour data, generalized linear models (GLMs) with non-normal distributions or non-parametric tests were used for the looming behaviour analysis. All run count data were analysed using a GLM with a Poisson distribution. All tail-rattle count data were analysed using a GLM with a quasi-Poisson distribution because the data were over-dispersed. Freeze, hide and ambulation data did not fit well to regression models, so non-parametric Kruskal–Wallis tests with Dunn's correction were performed. Percentages of mice and binary data were analysed with a Fisher's exact test. All statistical tests were two-tailed. Significance levels are indicated as follows: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$. For representative images, similar results were obtained in at least three independent trials.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

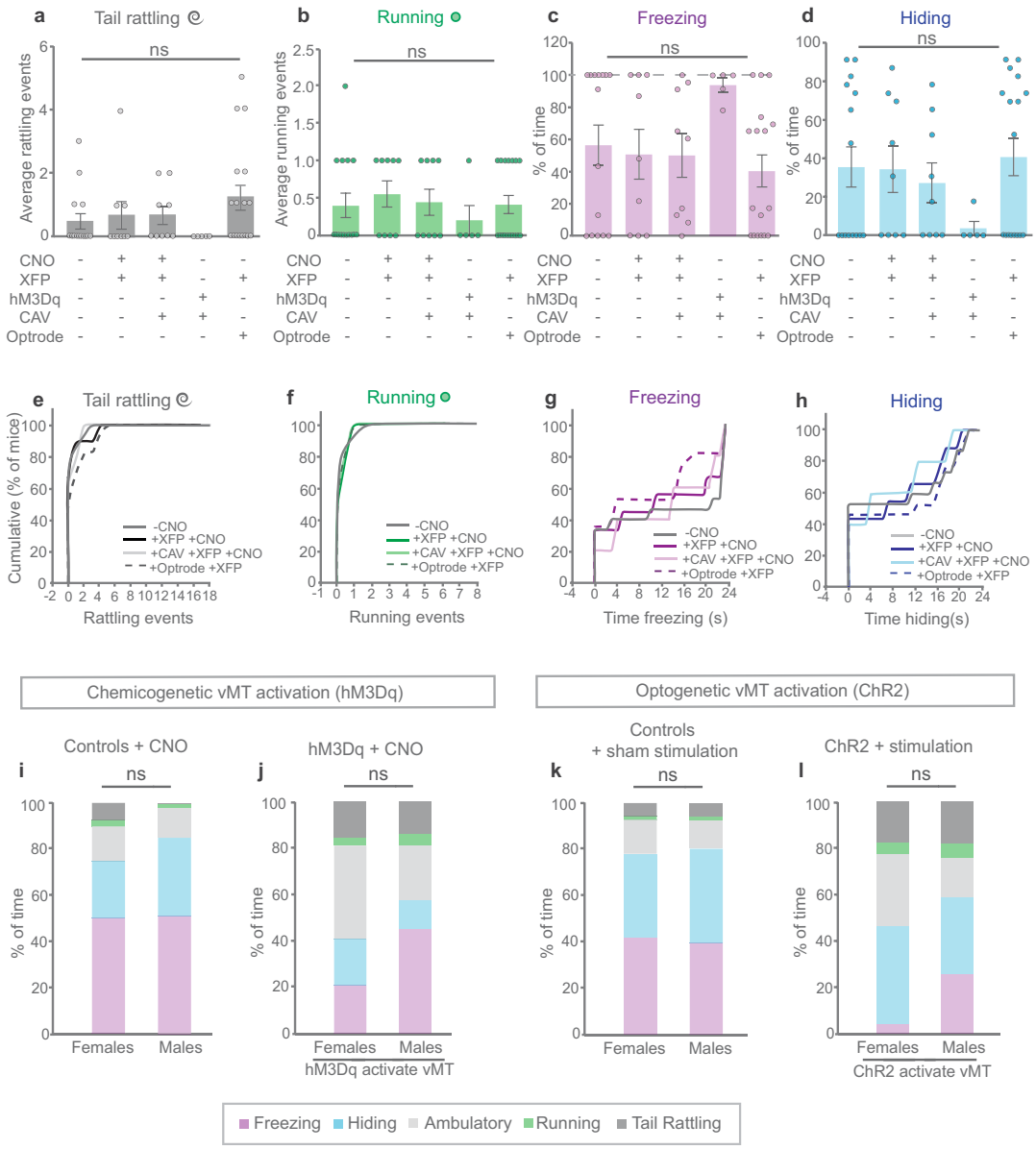
Data availability. The data that support the finding of this study are available from the corresponding authors upon reasonable request.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



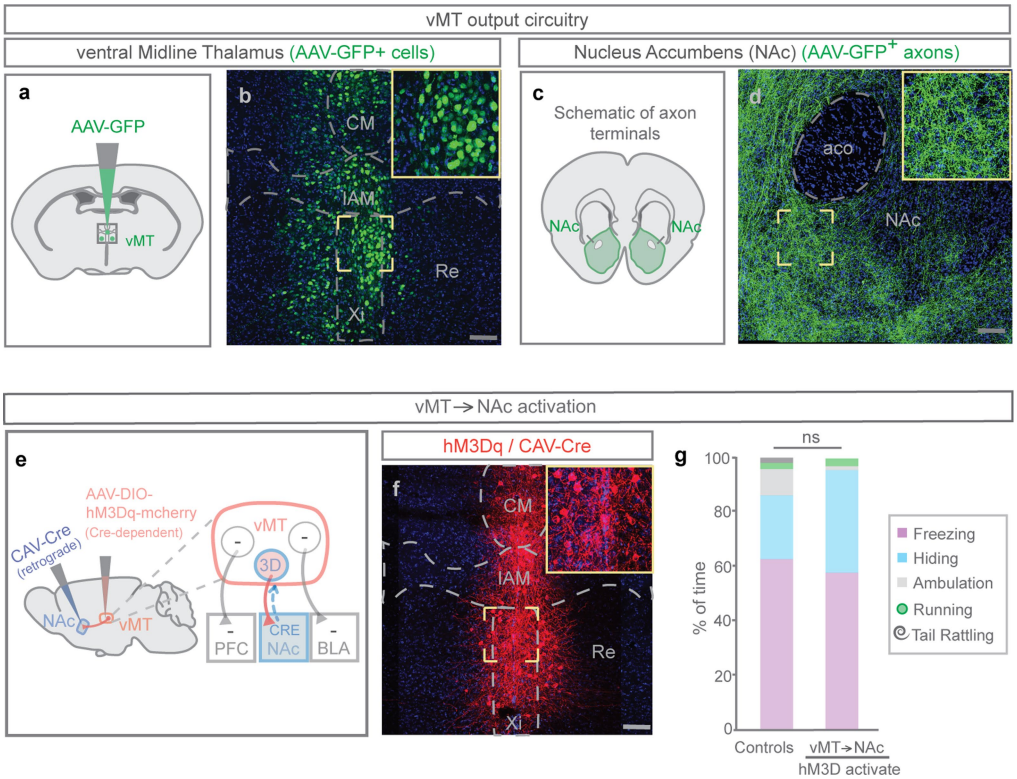
Extended Data Fig. 1 | hM3Dq activates vMT, whereas hM4Di inactivates vMT. Refers to Figs. 1 and 2. **a**, Timeline of the c-Fos induction protocol. **b**–**d**, Recently active c-Fos^{+} neurons (green) in the vMT of mice that were exposed to the looming stimulus, with XFP (**b**), hM4Di (**c**), or hM3Dq (**d**) injections (red) into the vMT and CNO delivered intraperitoneally. **e**, After CNO delivery, hM3Dq increased whereas hM4Di decreased the number of c-Fos^{+} cells in the vMT relative to XFP controls. Scale bars, 100 μm. Data are mean ± s.e.m. $*P < 0.05$, $***P < 0.001$. See Supplementary Table 1 for statistical analysis and sample sizes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



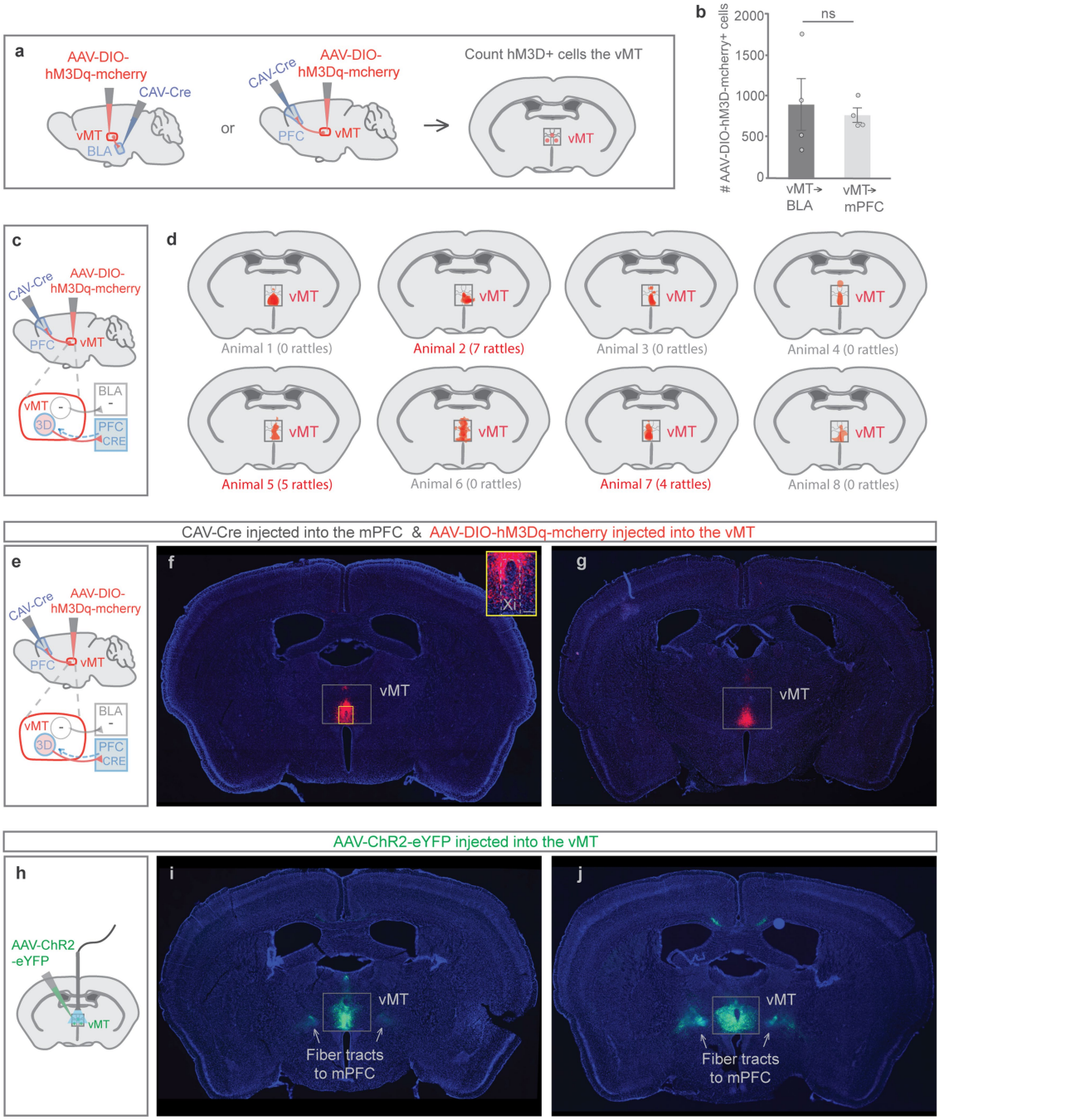
Extended Data Fig. 2 | Control groups do not differ in their defensive responses to looming threat. Refers to Figs. 2–4. **a**–**h**, Comparison of different control mice reveals no significant differences across any of the behaviours performed in response to the looming threat: tail rattling (**a**, **e**), running (**b**, **f**), freezing (**c**, **g**), or hiding (**d**, **h**). Controls include: mice with no treatments; mice with AAV-XFP and CNO; mice with CAV-Cre, AAV-DIO-XFP and CNO; mice with CAV-Cre and AAV-DIO-hM3Dq but without CNO; and mice with AAV-XFP, optrode implant and sham stimulation. **i**–**l**, Comparison of male and female control mice (**i**, **k**) and mice with vMT activation (**j**, **l**) reveals no significant sex differences across any of the behaviours performed in response to the looming threat. Notably, both male and female mice tail-rattle in response to the looming threat. Data are mean ± s.e.m.; NS, not significant. See Supplementary Table 1 for statistical analysis and sample sizes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



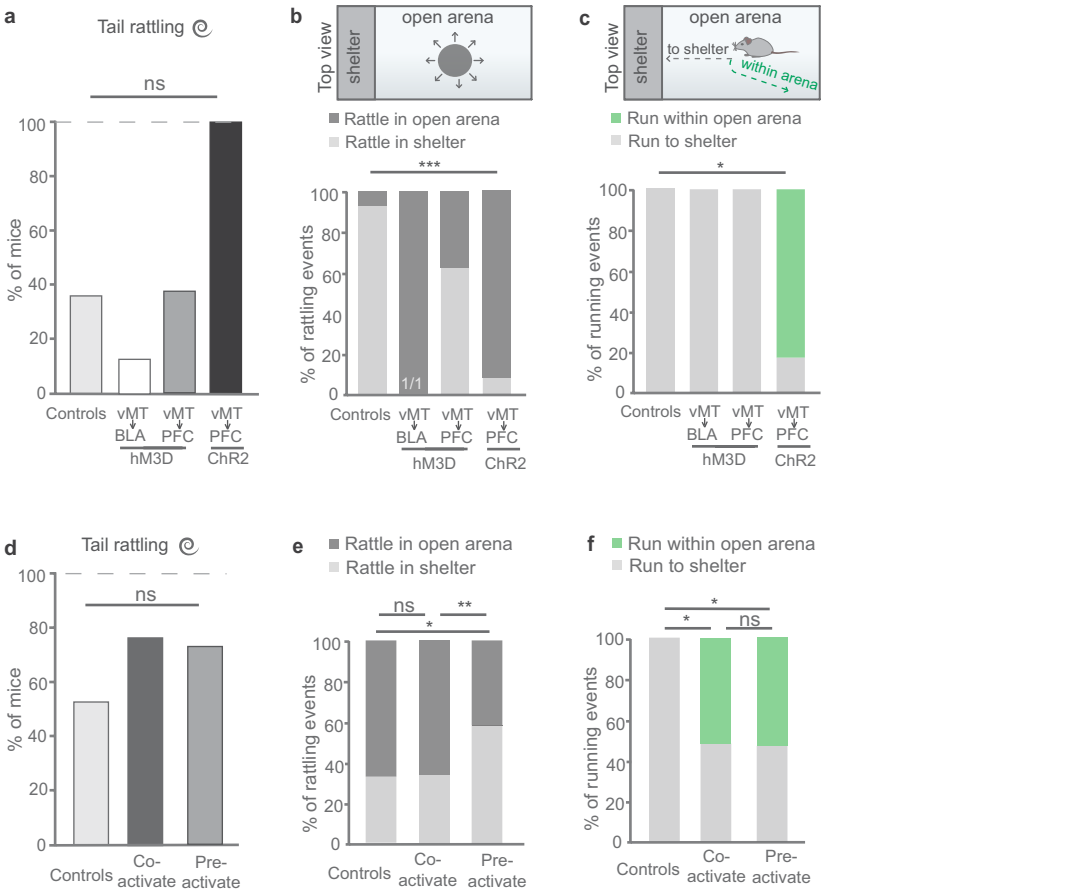
Extended Data Fig. 3 | Activating vMT→NAc does not increase tail rattling or saliency-enhancing behaviours. Refers to Fig. 3. **a**–**d**, Mice were injected with AAV-GFP in the vMT ($n = 17$ mice; **a**, **b**) and axons were observed in the NAc (**c**, **d**). Representative image of GFP^+ neurons in the VMT (**b**) and GFP^+ axons in the NAc (**d**). **e**, To activate vMT neurons that project to the NAc, CAV-Cre was injected into the NAc and Cre-dependent hM3Dq was injected into the vMT. **f**, Representative image of $hM3Dq^+$ neurons in the vMT that project to the NAc. **g**, Activating the vMT→NAc pathway did not significantly change the behavioural responses to looming as compared to controls. Scale bars, 100 μm. Data are mean ± s.e.m.; NS, not significant. See Supplementary Table 1 for statistical analysis and sample sizes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



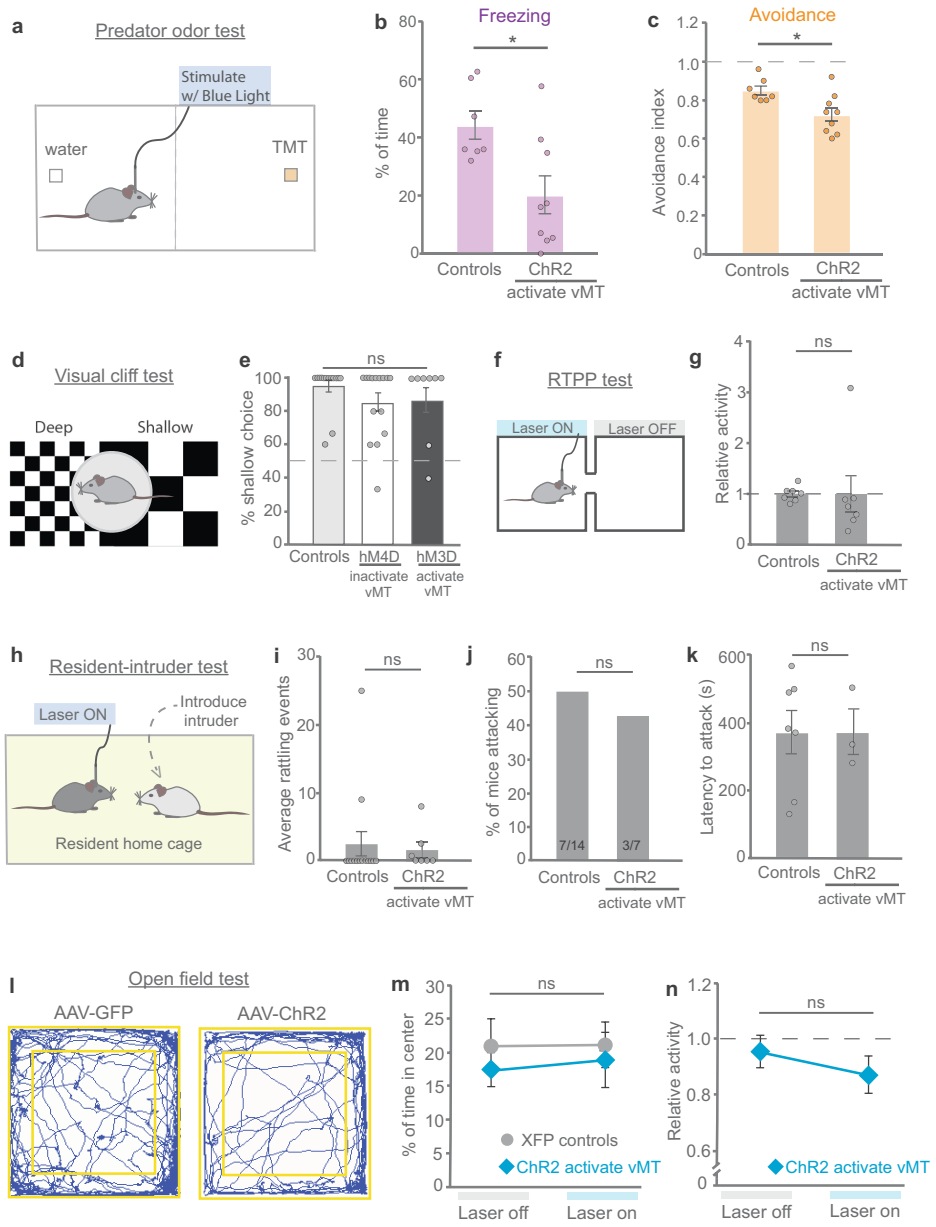
Extended Data Fig. 4 | Viral targeting and number of cells infected in the vMT. Refers to Figs. 3 and 4. **a**, To activate vMT neurons that project to the mPFC or the BLA, CAV-Cre was injected into the mPFC or the BLA and Cre-dependent hM3D was injected into the vMT. **b**, The average number of infected hM3d–mCherry$^+$ vMT cells did not differ between the vMT→BLA and the vMT→mPFC pathway activation groups. **c**, **e**, Locations of injections to activate the vMT→mPFC pathway. **d**, Relative expression of hM3D in vMT→mPFC cells does not scale with tail-rattling behaviour. **f**, **g**, Representative images of hM3dq–mCherry/Cre$^+$ neurons (red) in the vMT that project to the mPFC. **h**, Mice were injected with AAV-ChR2 in the vMT to activate the vMT. **i**, **j**, Representative images of ChR2–eYFP$^+$ neurons (green) in the vMT and the fibre tracts composed of vMT axons that project to the mPFC. Data are mean ± s.e.m.; NS, not significant. See Supplementary Table 1 for statistical analysis and sample sizes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



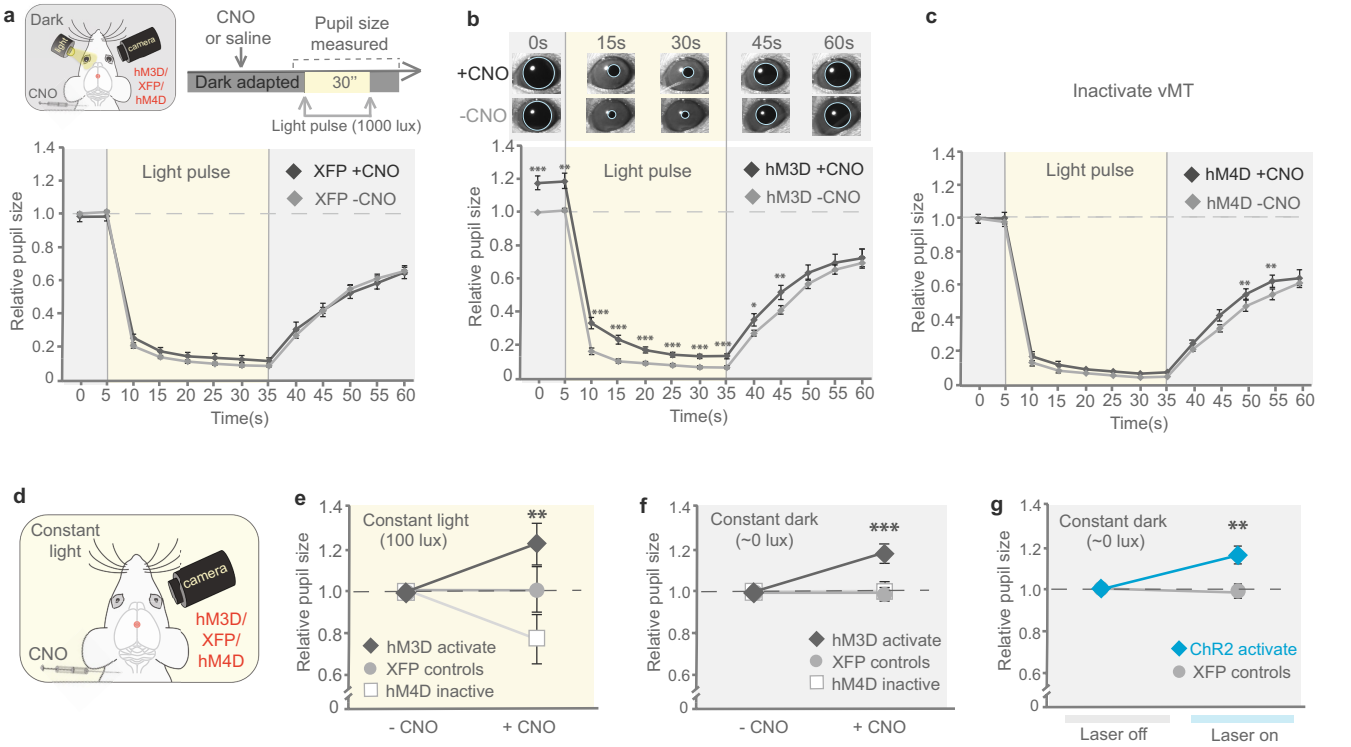
Extended Data Fig. 5 | vMT activation results in saliency-enhancing behaviours performed in the open arena. Refers to Figs. 3 and 4. **a**, Percentage of mice tail-rattling in response to looming after vMT→PFC or vMT→BLA activation. All of the mice with vMT→mPFC optogenetic stimulation displayed tail-rattling behaviour. **b**, **c**, Percentage of tail-rattling (**b**) or running (**c**) events performed in the open arena as opposed to in the shelter. Mice with vMT→ mPFC optogenetic stimulation perform most tail-rattling and running events in the open. **d**, Optogenetic activation of the vMT results in most mice tail-rattling. **e**, Mice with vMT optogenetic stimulation perform most tail-rattling events in the open arena as opposed to in the shelter. **f**, Mice with vMT optogenetic stimulation perform most running events within the open arena as opposed to towards the shelter. Data are mean ± s.e.m. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; NS, not significant. See Supplementary Table 1 for statistical analysis and sample sizes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



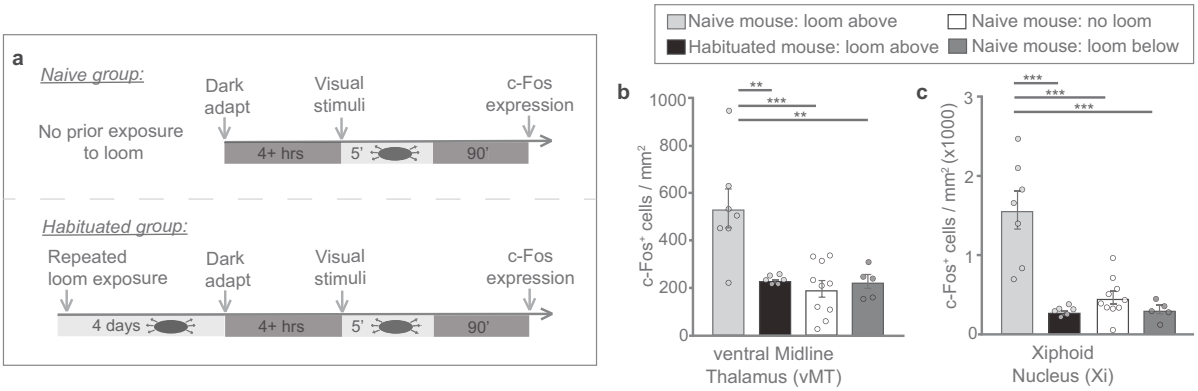
Extended Data Fig. 6 | vMT activation does not change locomotion, aggression or anxiety, but does result in less freezing in response to a predator odour. Refers to Figs. 4 and 5. **a**, Mice were injected with AAV-ChR2 in the vMT in order to activate the vMT during presentation of the predator-odour threat. **b**, Activating the vMT ($n = 9$ mice) decreased freezing as compared to controls with sham stimulation ($n = 7$ mice). **c**, Activating the vMT decreased avoidance as compared to controls with sham stimulation (avoidance index $= (P - 50)/50$, where P is the percentage of time the mouse spent on the side of the arena away from the odour). **d**, The visual cliff test. **e**, Activating the vMT (hM3D + CNO, $n = 15$ mice) or inactivating the vMT (hM4D + CNO, $n = 9$) during the visual cliff test did not change the number of times in which the mouse chose the shallow side as compared to control mice ($n = 9$ mice, $P < 0.05$). **f**, The RTPP test. **g**, Activating the vMT (ChR2, $n = 7$ mice) did not change the relative activity (distance covered) in the RTPP test as compared to control mice ($n = 6$ mice). **h**, Mice were tested on the resident-intruder test for aggressive behaviours. **i**, Activating the vMT (ChR2, $n = 7$ mice) did not change the average number of tail-rattling events in the resident-intruder test as compared to control mice ($n = 14$ mice). **j**, **k**, Activating the vMT (ChR2, $n = 7$ mice) did not change the percentage of mice attacking (**j**) or the latency to attack (**k**) in the resident-intruder test as compared to control mice ($n = 14$ mice). **l**, Mice were subjected to an open field test to analyse anxiety-related behaviour; representative tracing of a control mouse (AAV-GFP, left) and a mouse with vMT activation (AAV-ChR2, right) in the open field test. **m**, Activating the vMT (ChR2, $n = 10$ mice) did not change the percentage of time in the centre of the open field as compared to control mice ($n = 10$ mice). **n**, Activating the vMT (ChR2, $n = 10$ mice) did not change the relative activity (distance covered) in the open field test as compared to control mice ($n = 10$ mice). Data are mean ± s.e.m. $^*P < 0.05$; NS, not significant. See Supplementary Table 1 for statistical analysis and sample sizes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



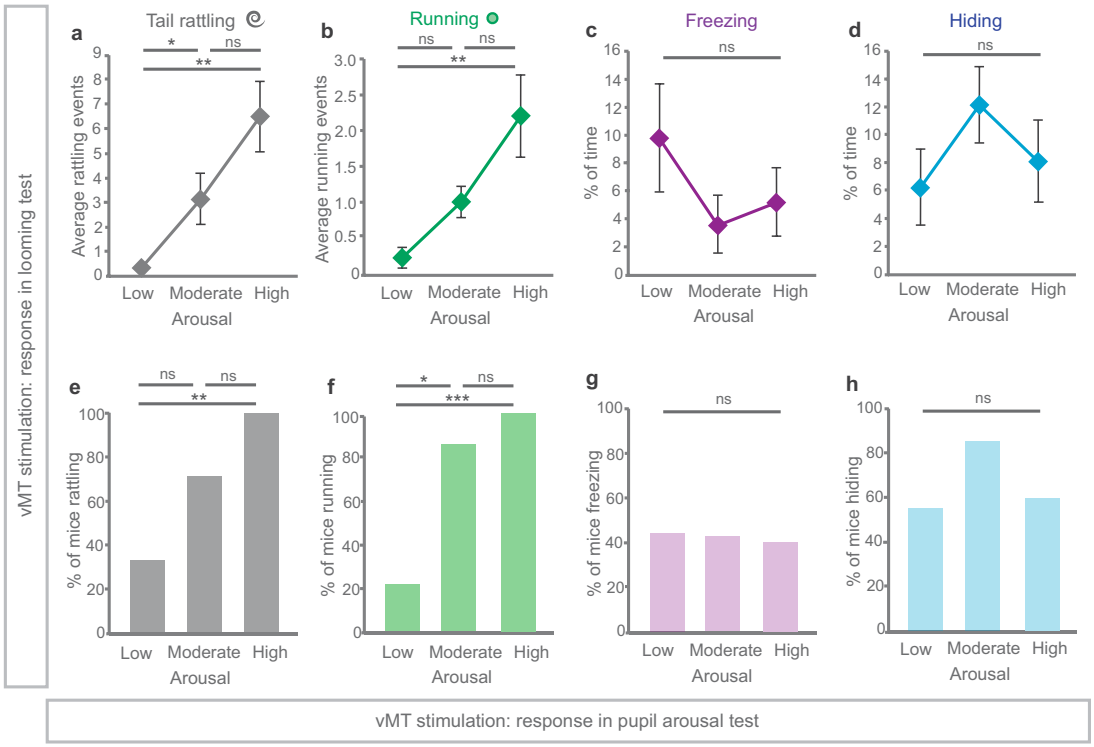
Extended Data Fig. 7 | vMT activation results in increased arousal. Refers to Fig. 5. **a**, Schematic of the experimental set-up to analyse light-driven pupillary responses in mice. Treating control mice with CNO did not change pupil size across all light levels. **b**, Chronic vMT activation by chemicogenetic methods (+CNO) significantly increases pupil size across all light levels as compared to the same mice without activation (−CNO). **c**, Chronic vMT inactivation (+CNO) did not the change pupil size across all light levels as compared to the same mice without activation (−CNO). **d**, Schematic of experimental set-up to analyse arousal-driven pupillary responses in mice with vMT activation, mice with vMT inactivation and control mice. The pupil was measured in constant light (100 lx) conditions in mice with and without CNO. **e**, After CNO delivery, mice with hM3Dq had a significant increase in relative pupil size. Mice with XFP and hM4Di did not have a significant change in pupil size after administration of CNO. **f**, In constant dark conditions, vMT activation significantly increased pupil size, whereas vMT inactivation did not change pupil size as compared to control mice with CNO. **g**, In constant dark conditions, optogentic activation of the vMT significantly increased pupil size ($n = 11$ mice) compared to control mice ($n = 12$ mice). Data are mean ± s.e.m. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; NS, not significant. See Supplementary Table 1 for statistical analysis and sample sizes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



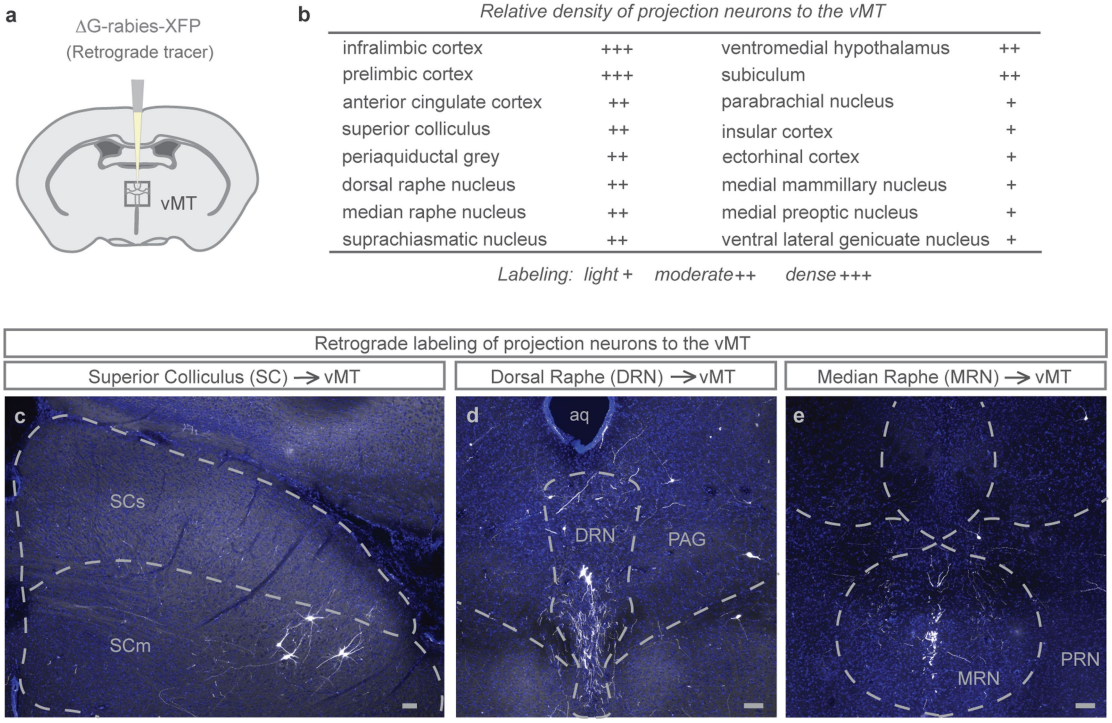
Extended Data Fig. 8 | Looming stimuli induce vMT activation in naive, but not experienced mice habituated to looms. Refers to Fig. 6. **a**, Schematic of the experimental protocol. **b**, **c**, Quantification of c-Fos$^+$ cells in the vMT of naive or pre-exposed mice that experienced looming from above revealed a significant decrease in c-Fos$^+$ cells in the vMT (**b**) and Xi (**c**) after looming from above in pre-exposed mice that are habituated to the looms ($n = 6$ mice) as compared to naive mice ($n = 7$ mice). Data are mean ± s.e.m. $^{**}P < 0.01$; $^{***}P < 0.001$. Supplementary Table 1 for statistical analysis and sample sizes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Tail rattling and running scales with arousal. Refers to Figs. 4 and 5. **a**, **b**, Mice in which ChR2 had been injected into the vMT were divided into three groups based on the extent to which vMT activation increased arousal responses: low, moderate or high ($n = 7$, $n = 9$ and $n = 10$ mice, respectively). Mice with high arousal had significantly more tail-rattling (**a**) and running (**b**) events in response to looming, as compared to mice with low arousal ($P < 0.001$ and $P < 0.001$). **c**, **d**, Mice with high arousal spent similar amounts of time freezing (**c**) and hiding (**d**) in response to looming, as compared to mice with low arousal. **e**, **f**, 100% of mice with high arousal tail-rattled (**e**) and ran (**f**) in response to looming, whereas only 33% and 20% of mice with low arousal tail-rattled and ran, respectively. **g**, **h**, Similar numbers of mice with high arousal froze (**g**) and hid (**h**) in response to looming as compared to mice with low arousal. Data are mean ± s.e.m. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; NS, not significant. See Supplementary Table 1 for statistical analysis and sample sizes.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Hindbrain inputs to the vMT. a, Schematic illustration of ΔG-rabies-XFP injection into the vMT to map vMT inputs. **b**, Quantification of the relative density of projection neurons to the vMT ($n = 12$ mice). **c**–**e**, Representative images showing the expression of ΔG-rabies-XFP in transynaptically labelled cells in the superior colliculus (**c**), dorsal raphe (**d**), periaqueductal grey (**d**), and median raphe (**e**). DRN, dorsal raphe; MRN, median raphe; PAG, periaqueductal grey; PRN, pontine reticular nucleus; SCm, superior colliculus, motor; SCs, superior colliculus, sensory. Scale bars, 100 μm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0062-x

Cryo–EM structure of substrate–bound human telomerase holoenzyme

Thi Hoang Duong Nguyen[1,2,3,4], Jane Tam[1], Robert A. Wu[1,6], Basil J. Greber[2,3], Daniel Toso[2], Eva Nogales[1,2,3,5]* & Kathleen Collins[1,2]*

The enzyme telomerase adds telomeric repeats to chromosome ends to balance the loss of telomeres during genome replication. Telomerase regulation has been implicated in cancer, other human diseases, and ageing, but progress towards clinical manipulation of telomerase has been hampered by the lack of structural data. Here we present the cryo–electron microscopy structure of the substrate–bound human telomerase holoenzyme at subnanometre resolution, showing two flexibly RNA–tethered lobes: the catalytic core with telomerase reverse transcriptase (TERT) and conserved motifs of telomerase RNA (hTR), and an H/ACA ribonucleoprotein (RNP). In the catalytic core, RNA encircles TERT, adopting a well–ordered tertiary structure with surprisingly limited protein–RNA interactions. The H/ACA RNP lobe comprises two sets of heterotetrameric H/ACA proteins and one Cajal body protein, TCAB1, representing a pioneering structure of a large eukaryotic family of ribosome and spliceosome biogenesis factors. Our findings provide a structural framework for understanding human telomerase disease mutations and represent an important step towards telomerase–related clinical therapeutics.

Linear eukaryotic chromosomes are capped by telomeric repeat tracts that recruit the telomere-binding proteins that are essential to distinguish telomeres from DNA breaks and thus to avoid telomere end-resection and interchromosomal fusions[1,2]. Owing to inherently incomplete genome replication, telomeres are progressively shortened in each cell cycle[3]. Critically short telomeres result in genome instability and cell death[2,4,5]. To compensate for this sequence loss, a specialized reverse transcriptase, telomerase, adds telomeric repeats to the chromosome 3′ end, using TERT and an integral telomerase RNA subunit (TER) with an internal template for repeat synthesis[6]. Human telomerase activation in embryogenesis and its repression in somatic tissues govern cellular renewal capacity, with telomerase deficiency imposing haematopoietic and epithelial failures and aberrant telomerase activation enabling tumorigenesis[7].

Despite its medical relevance, efforts to design drugs that target telomerase have been hampered by limited structural information. Endogenous human telomerase holoenzyme is scarce. Most studies of human telomerase use cellular overexpression of limiting TER (hTR in humans) and TERT[8]. Two hTR domains are critical for activity[9,10]: a domain with the template and adjacent pseudoknot (t/PK) and the conserved regions 4 and 5 (CR4/5), consisting of a branched junction of stems P5 and P6, and the activity-critical stem-loop P6.1 (Fig. 1a). Together with TERT, these domains are sufficient to reconstitute telomerase activity with the repeat addition processivity characteristic of the cellular holoenzyme[11]. TERT is composed of four domains: the N-terminal (TEN) domain, the high-affinity RNA-binding domain (TRBD), the reverse transcriptase domain (RT), and a C-terminal extension (CTE) analogous to a polymerase thumb[11] (Fig. 1a). The high-resolution structure of a streamlined TERT from the flour beetle *Tribolium castaneum*, which lacks the TEN domain, established that the latter three domains form a ring, referred to as the TERT ring[12]. The TERT ring alone can bind a template–primer duplex[13], and, accordingly, the human TERT ring has been shown to support single-repeat synthesis[14]. Purified or overexpressed hTR and TERT form heterogeneous complexes, only some of which are catalytically active[15]. It remains unclear whether the active enzyme requires an hTR and TERT dimer[16,17] or only a monomer of each subunit[15,18]. The landmark 9 Å cryo-electron microscopy (cryo-EM) structure of the ciliate *Tetrahymena thermophila* telomerase holoenzyme revealed its overall architecture, including a monomeric TERT–TER catalytic core[19]. However, the subunit composition of the *Tetrahymena* telomerase holoenzyme is distinct from its human counterpart, and available structural information on human telomerase holoenzyme is limited to a 30 Å negative-stain electron microscopy reconstruction[17]. Therefore, both the structure and the composition of the human telomerase holoenzyme remain poorly characterized.

Telomerase assembly initiates in human cells with nascent hTR binding to H/ACA proteins and their chaperones[20]. Human H/ACA proteins are shared by telomerase and the small nucleolar (sno) and small Cajal body (sca) RNP families, which catalyse ribosomal and spliceosomal RNA modifications, respectively[21,22]. Eukaryotic H/ACA snoRNPs contain the pseudouridine synthase dyskerin (Cbf5), NOP10, NHP2, and GAR1, assembled on each of two hairpin stems; and scaRNPs also contain TCAB1/WDR79 bound to loop CAB-box motifs. The hTR H/ACA 3′ hairpin loop has a CAB box but the 5′ hairpin instead extends into CR4/5 (Fig. 1a). All eukaryotic H/ACA RNPs, including telomerase, require both hairpins of the H/ACA motif for RNP biogenesis in cells[22], unlike their single-hairpin archaeal counterparts[21], but it remains unclear whether the active human telomerase holoenzyme has one or two sets of H/ACA proteins[18], only dyskerin[23], or only NOP10 and dyskerin[17]. Syndromes of human telomerase deficiency have been traced to mutations in dyskerin, NOP10, NHP2 and their chaperone NAF1[5]. While the structure of the related single-hairpin archaeal RNP is known[24], there is no high-resolution structure of any complete eukaryotic H/ACA RNP.

Here, to address the structure and composition of human telomerase holoenzyme, we optimized active telomerase purification and used cryo-EM to determine its structure at subnanometre resolution, which

[1]Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA. [2]California Institute for Quantitative Biology, University of California, Berkeley, CA, USA. [3]Molecular Biophysics and Integrative Bio-Imaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [4]Miller Institute for Basic Research in Science, University of California, Berkeley, CA, USA. [5]Howard Hughes Medical Institute, University of California, Berkeley, CA, USA. [6]Present address: Harvard Medical School, Boston, MA, USA. *e-mail: kcollins@berkeley.edu; enogales@lbl.gov

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

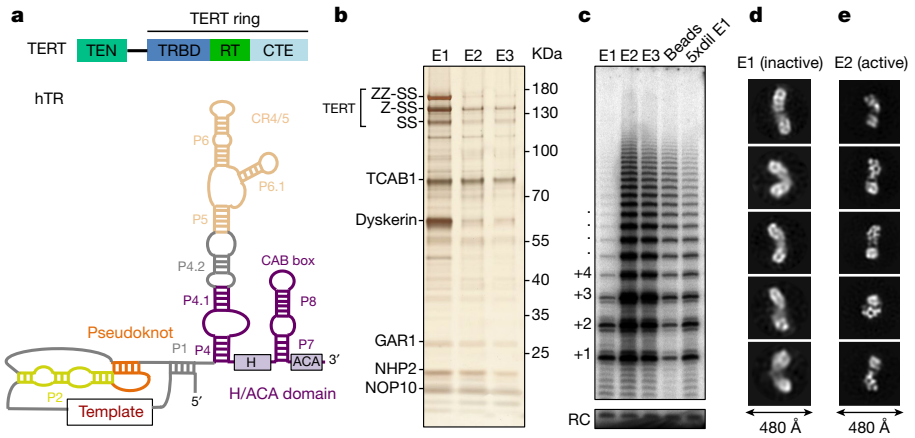


Fig. 1 | Telomerase holoenzyme reconstitution and characterization. **a**, Domain architecture of TERT (top) and secondary structure of hTR (bottom). Paired stems (P) are shown compacted in length. **b**, Silver-stained SDS–PAGE gel showing fractionation of purified, heterogeneously active TERT RNP. Proteins were detected by mass spectrometry and immunoblotting (Extended Data Fig. 1a, Extended Data Table 1). **c**, In vitro telomerase assays performed on the three elution fractions shown in **b**, the beads after elution, and fivefold diluted E1. A 12-nucleotide (nt) oligonucleotide was used as recovery control (RC). Processive repeat additions are numbered or indicated by dots. **d**, **e**, Representative negative-stain 2D class averages for fractions E1 and E2, respectively. We also observed a small fraction of E2-like particles in E1 (data not shown), while E2 contained predominantly the particles shown in **e**. Experiments in **b** and **c** were performed three times and twice, respectively, with similar results. For gel source data, see Supplementary Fig. 1.

allowed us to fit available high-resolution structures and homology models of all components. Our structure reveals in detail the organization of ten protein subunits and hTR, illuminating the structural basis of hTR motif functions and human disease mutations that cause telomerase deficiency.

Activities of distinct structural states

We reconstituted human telomerase holoenzyme by transient transfection of HEK 293T cells with vectors expressing TERT and hTR[8]. After affinity purification steps to enrich for hTR and TERT, the predominant additional proteins in the sample were TCAB1 and the H/ACA proteins (dyskerin, NOP10, NHP2 and GAR1), each of which was assigned by immunoblotting and, with the exception of glycine/arginine-rich GAR1, also confirmed by mass spectrometry (Fig. 1b, Extended Data Fig. 1a, Extended Data Table 1). Furthermore, protein identities were confirmed by subunit drop-out purifications (see below). Direct primer-extension assays of eluted fractions showed that the amount of TERT did not fully correlate with the amount of telomerase activity; also, dilution of the less-active fraction E1 increased rather than decreased activity (Fig. 1c). Visualization by negative-stain electron microscopy of E1 and E2, which we refer to as inactive and active fractions, respectively, revealed morphologically distinct populations of particles. Particles from the inactive fraction were more elongated (typically 400 Å or more in length) and more heterogeneous than those in the active fraction (Fig. 1d, e). Particles from the active fraction consisted of two asymmetric lobes, nearly 300 Å in length (Fig. 1e). Previous studies have shown that cellular human telomerase reconstitution yields complexes both monomeric and dimeric for TERT[15,17], explaining the observed particle heterogeneity. For cryo-EM studies, we modified our purification protocol to recover only active fractions (Extended Data Fig. 1b, c), added a DNA substrate ($T_{14}AG_3$) during purification, and confirmed the presence of the bound substrate in the purified RNP (Extended Data Fig. 1c). We also demonstrated that our tagged TERT was active in vivo (Extended Data Fig. 2a–e).

Architecture and flexibility of telomerase

We prepared ΔTCAB1 telomerase with an hTR CAB box mutant that disrupted TCAB1 binding[25,26] and separately prepared TERT–hTRmin[27,28], in which hTRmin contains only the activity-essential t/PK and CR4/5 domains of hTR. By comparing their negative-stain 2D class averages to those of the intact holoenzyme, we established the particle regions corresponding to TCAB1, the H/ACA RNA motif and proteins, and TERT with the two activity-essential domains of hTR (Extended Data Fig. 1d–h). This revealed that one lobe of the holoenzyme contained TERT and its associated hTR motifs (the catalytic core) and the other contained the H/ACA RNP components and TCAB1 (the H/ACA lobe). We next collected cryo-EM data to determine the structure of the substrate-bound human telomerase holoenzyme (Extended Data Fig. 3a, b, Extended Data Table 2a). Initial image processing revealed extensive conformational flexibility between the two lobes, as well as within each lobe, which limited the quality of the overall reconstruction (Extended Data Figs. 3c, 4a). We performed focused classification with signal subtraction[29] to obtain more homogeneous subsets of particles for each lobe separately. This procedure yielded reconstructions at 7.7 Å and 8.2 Å resolution for the catalytic core and H/ACA lobes, respectively (Extended Data Figs. 3c, 4b–e). At these resolutions, densities for protein α-helices, β-barrels, and double-stranded RNA are discernible, allowing us to unambiguously fit available protein and RNA structures or homology models, and ultimately to account for the majority of the density, except for flexible regions of single-stranded RNA and protein extensions (Fig. 2a, b, Extended Data Fig. 4f). We also obtained a reconstruction of the entire holoenzyme at 10.2 Å resolution (Extended Data Fig. 3c).

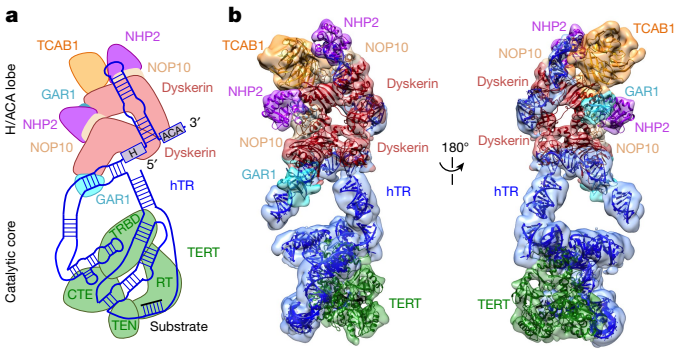


Fig. 2 | Cryo-EM structure of the substrate-bound human telomerase holoenzyme. a, Schematic of subunit arrangements. **b**, Front (left) and back (right) views of the cryo-EM reconstructions for the H/ACA lobe at 8.2 Å and the catalytic core at 7.7 Å, with fitted subunits colour-coded as indicated.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

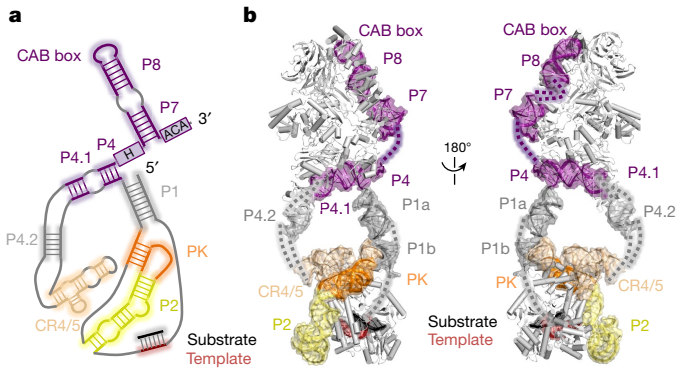


Fig. 3 | hTR structure in human telomerase holoenzyme. a, Schematic of secondary structure of hTR, with domains arranged based on the cryo-EM reconstruction. Regions modelled are highlighted with soft-coloured backgrounds. **b**, Front (left) and back (right) views of the structure of hTR highlighted within the human telomerase holoenzyme structure, with domains colour-coded as in **a** and unmodelled connections shown as dashes lines.

Subunit distribution across the two lobes

The bilobal human telomerase includes one TERT molecule and one hTR (Fig. 2a, b), in agreement with suggestions of an evolutionarily conserved single-TERT catalytic core for telomerases from ciliate and yeast model systems[30,31] to human[15,18,32]. The complex also contains one TCAB1 and two sets of the four mature H/ACA RNP proteins. The previous human telomerase negative-stain electron microscopy reconstruction also had a bilobal architecture, interpreted as dimeric TERT, although gold-labelling experiments indicated the presence of a mixture of RNPs containing both one (50%) and two (36.5%) TERT molecules[17]. Incomplete separation of active from inactive TERT assemblies may account for the mixture of complexes in the previous study. Our finding of separate lobes for the catalytic core and the H/ACA RNP gives a structural explanation for why telomerase holoenzyme RNP biogenesis strategies have such plasticity across eukaryotes[11].

The extended architecture of hTR scaffolds the overall subunit organization across both lobes (Fig. 3a, b). The catalytic core and H/ACA lobes are bridged by stems P1a and P4.2 (Fig. 2a, b, 3b). The crystal structure of *Tribolium* TERT in complex with an RNA–DNA duplex[13] could be readily fitted within the catalytic core density (Fig. 4a, Extended Data Figs. 5c, d, 6b). As the *Oryzias latipes* (medaka) TRBD structure[33] includes vertebrate TERT motifs not present in the streamlined *Tribolium* TERT (Extended Data Fig. 7e), we removed extensions from the medaka TRBD structure that did not fit our density and used this model to replace the *Tribolium* TRBD (Extended Data Fig. 5b). Adjacent to the TERT ring, there is globular density with approximately the dimensions of the *Tetrahymena* TEN domain crystal structure[34] (Fig. 4b, Extended Data Fig. 5a). The *Tetrahymena* and human TEN domains share only 20% sequence identity and have differences in predicted secondary structure[35]; therefore, we removed regions of the former that are likely to be less structurally congruent with the human structure for tentative fitting into our map (Extended Data Fig. 6a). The NMR structure for the full human PK[36] and the crystal structure of the medaka CR4/5 in the conformation bound to TRBD[33] were fitted in one side of the catalytic core (Figs. 2b, 3a, b, 4a, Extended Data Fig. 5h). The PK is connected to the template region located in the TERT active site by a bent helical RNA density that we attributed to the sequential junction and stem elements of P2 (P2b–J2a/b–P2a–J2a/a.1–P2a.1)[10,37,38] (Figs. 3a, 4a).

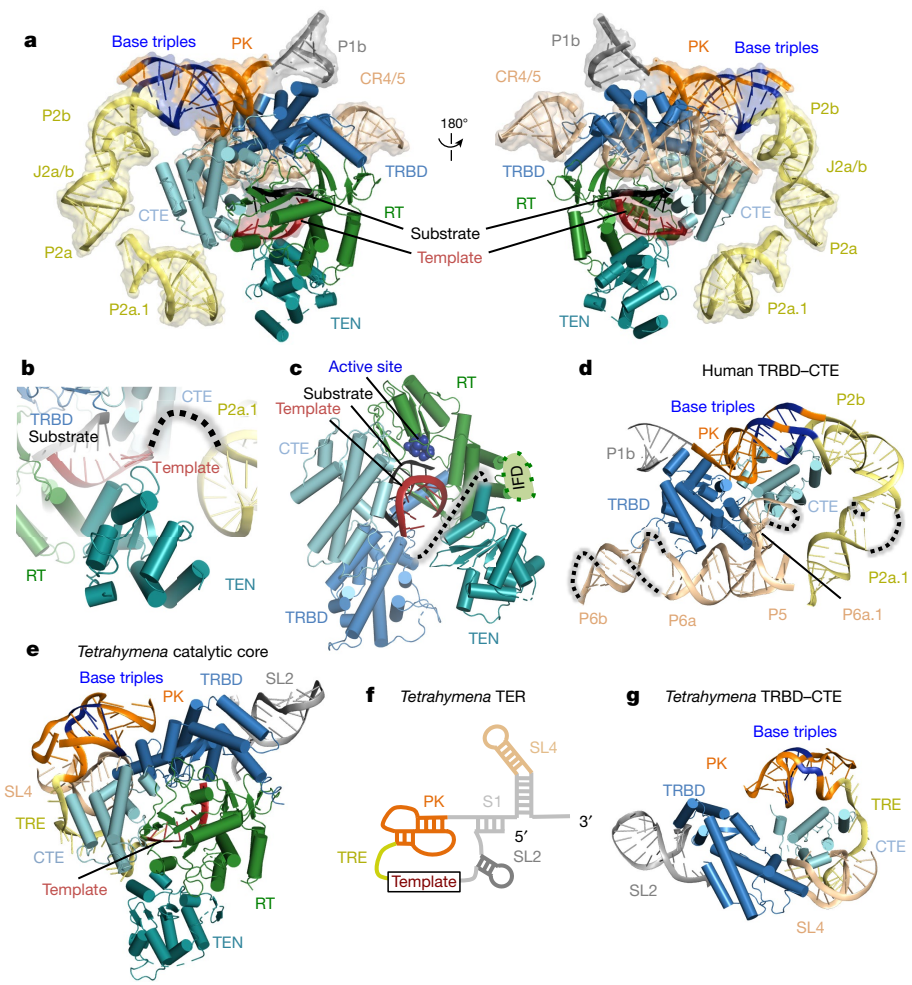


Fig. 4 | Structure of the catalytic core. a, Front (left) and back (right) views of the catalytic core. Domains of human TERT and hTR are colour-coded as in Fig. 1a. hTR is in semi-transparent space-filling representation, and base triples within the PK are indicated in blue. **b**, Route of single-stranded RNA connecting hTR P2a.1 and the template, which threads near the TEN domain. **c**, TERT interactions with the substrate–template duplex. Active site residues are shown as blue spheres. The human TERT IFD insertion missing in the *Tribolium* TERT is indicated between helices α10 and α11 of the *Tribolium* TERT[13] (Extended Data Fig. 6b). A dotted line depicts the linker connection of TEN and TRBD domains. **d**, The human TRBD and CTE domains are encircled by hTR P2 (yellow) and PK (orange) and CR4/5 (wheat) domains. **e**, **f**, Architecture of *Tetrahymena* TERT domains with TER motif colours as indicated[19]: PK, orange; template recognition element (TRE), yellow; SL2, dark grey; SL4, wheat. Base triples within the PK are indicated in blue. For clarity, domains of TER coloured light grey in **f** are not shown in the structure in **e**. **g**, The *Tetrahymena* TRBD and CTE domains are encircled by t/PK (orange), TRE (yellow), SL2 (grey) and part of SL4 (wheat). Base triples within the PK are indicated in blue.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

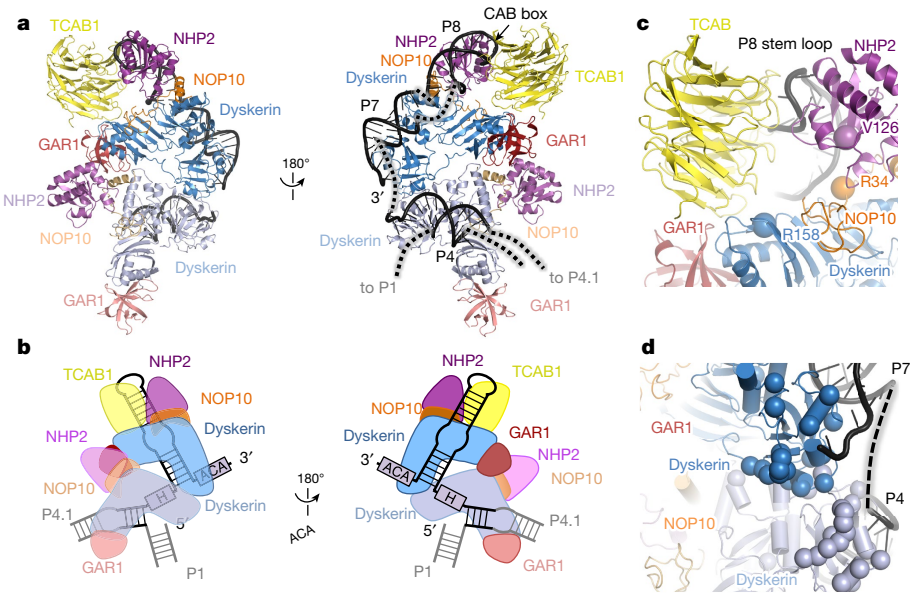


Fig. 5 | The H/ACA RNP and disease mutations. a, Front (left) and back (right) views of the H/ACA lobe within the human telomerase structure, with colour-coded subunits: TCAB1, yellow; dyskerin, blue; GAR1, red; NOP10, orange; NHP2, magenta; hTR, black. To distinguish between the two sets of H/ACA proteins, the first set bound to the 5′ hairpin (P4 stem) is in a lighter shade. **b**, Schematics of subunit arrangements within the H/ACA lobe. **c**, Interactions of TCAB1 with hTR and H/ACA proteins. The backbone amides of disease-associated mutations in dyskerin, NOP10 and NHP2 that mapped near TCAB1 and P8 interaction surfaces are shown as spheres. **d**, Mapping of a cluster of disease mutations (backbone amides shown as spheres) in dyskerin near the dyskerin–dyskerin and dyskerin–RNA interfaces.

We docked two copies of the crystal structure of an archaeal H/ACA RNP[24], each consisting of Cbf5 (orthologous to dyskerin), GAR1, NOP10 and L7Ae (structurally related to NHP2) and a single-hairpin RNA, into the H/ACA lobe (Figs. 2a–c, 5a, b, Extended Data Figs. 5e, f, 7f, 8a–d). The single-hairpin RNA was truncated to a 9-bp stem flanked by the ACA motif to mimic P4 and P7 stems flanked by box H and ACA, respectively, in hTR. The hTR segments bound by each archaeal RNP equivalent can be unambiguously assigned on the basis of their connectivity with the rest of the RNA densities. The P4–P4.1 stem is connected to the CR4/5 domain in the catalytic core via the P4.2 stem, and P7 is connected to the P8 stem-loop (Fig. 3a, b, 5a, b, Extended Data Fig. 5i). Density in the regions of P1, P4.1 and P4.2 was fitted with A-form RNA duplexes. TCAB1 harbours a WD40 domain, and its homology model fit uniquely at the top of the H/ACA lobe (Figs. 2b, 5a, b, Extended Data Figs. 5g, 7g), consistent with the missing density from the hTR CAB-box mutation (Extended Data Fig. 1h). This assignment of TCAB1 places it adjacent to an RNA-hairpin-like density that was fitted with the NMR structure of the P8 stem-loop[39].

The catalytic core

The t/PK and CR4/5 domains of hTR wrap around TERT to form the catalytic core (Fig. 4a). In turn, the human TERT ring encircles the fitted *Tribolium* TERT RNA–DNA bound to the TERT active site (Fig. 4a, c), which shifts the template position compared to that in the substrate-free *Tetrahymena* telomerase holoenzyme (Extended Data Fig. 7i). The majority of the TERT RNA binding surface is contributed by the TRBD and CTE, which are sandwiched between the PK and CR4/5 (Fig. 4d). The full human PK (P2b–P3) is connected to P2a by the J2a/b bulge, and P2a connects to the less conserved helical extension P2a.1 before the transition to the single-stranded 3′-template-flanking region (Figs. 1a, 3a, 4a). The t/PK forms a rigid, arc-like structure (Figs. 1a, 4a), as predicted by modelling using NMR structures of its subdomains[10,36,38]. The less conserved end of the PK P3 stem approaches the TRBD close to its interface with the CTE and co-axially stacks with P2b (Fig. 4d). The unexpectedly limited interaction between the hTR PK and TERT favours a model for indirect PK function in the co-folding of TERT and hTR[30,40]. The location of P2a.1 positions the immediately 5′ single-stranded 3′-template-flanking RNA such that it could thread between the TEN domain and the TERT ring (Fig. 4b, Extended Data Fig. 7c), similar to the path of the 3′-template-flanking TER in the *Tetrahymena* telomerase holoenzyme[9,10,19] (Fig. 4e), despite major changes in the remainder of the t/PK secondary and tertiary structure (Figs. 1a, 4f, Extended Data Fig. 7j). The proximity of the TEN domain to the template 3′ end further reinforces its proposed role in stabilization of RNA–DNA duplex at the template 3′ end[14,27]. The position of the *Tetrahymena* 5′ template-boundary element, stem-loop 2 (SL2), relative to the TRBD, is roughly similar to that of human P1b (Fig. 4d, g), consistent with their common function in 5′ template boundary determination[41].

The CR4/5 domain, which consists of P5, P6 and P6.1 stems connected as a three-way junction, adopts a similar conformation to that in the medaka CR4/5–TRBD crystal structure[33] (Fig. 4d, Extended Data Fig. 7d), except that human P6 is longer and composed of two stems, P6a and P6b. The P6a stem binds along TRBD (Fig. 4d), while the P6.1 hairpin emanates from the three-way junction almost perpendicularly to P6a, inserting into the interface between the TRBD and CTE. Clear helical density is observed for P6b, which makes no RNA–RNA or RNA–protein interactions. P5 co-axially stacks with P6a, but like P6b, it does not interact with TERT (Fig. 4d, Extended Data Fig. 7d). The P6.1-TERT interaction is critical for TERT association with hTR and telomerase catalytic activity[42,43]. The P6.1 apical loop was disordered in the TRBD–CR4/5 crystal structure[33] and, although our resolution is insufficient for de novo building, its density is visible in our cryo-EM map, appearing packed against the CTE (Extended Data Fig. 7d). Our structure suggests that interaction with the CTE stabilizes this loop, consistent with previous crosslinking studies[44]. In the *Tetrahymena* telomerase cryo-EM structure, the stem-terminus element (SL4) is bound at the interface between the TRBD and CTE of TERT[19], similar to P6.1 in our human telomerase structure (Fig. 4d, 4g). Whether or not they are related evolutionarily, P6.1 and SL4 serve a similar structural role in telomerase catalytic core architecture. Finally, the catalytic core has extra density that is likely to correspond to TERT and hTR regions that were not possible to model (Extended Data Fig. 4f), such as the 'insertion in fingers' (IFD) subdomain of the RT domain (residues 734–802) that is missing in *Tribolium* TERT (Extended Data Figs. 6b, 7a, b), and the non-conserved flexible linker between the TEN domain and the TERT ring (Fig. 4c).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

The H/ACA RNP and human disease mutations

Our structure offers the first glimpse of the subunit architecture of an intact eukaryotic H/ACA RNP, with two sets of H/ACA heterotetramer proteins that bind two RNA hairpins (Fig. 5a, b). As expected, the 3′ hairpin loop associates with the single holoenzyme TCAB1 subunit (Fig. 5a–c). We observe an interaction between the two H/ACA protein heterotetramers, predominantly mediated by the two dyskerin molecules (Fig. 5d, Extended Data Fig. 7h). The first set of H/ACA proteins contacts the 5′ hairpin P4 stem exclusively via dyskerin (Fig. 5a, b, Extended Data Fig. 5e). By contrast, the second set forms more extensive interactions with the 3′ hairpin P7 stem and P8 stem-loop. The P7 stem is bound by dyskerin, similarly to the 5′ hairpin P4, and continues to P8, which interacts with NOP10, NHP2 and TCAB1 (Fig. 5a, b, Extended Data Fig. 5f, i). The fact that binding of dyskerin at the base of the 5′ hairpin is sufficient to anchor the entire core heterotetramer accounts for the evolutionary divergence of the vertebrate TER 5′ H/ACA hairpin[45] and for the hTR 5′ hairpin tolerance of changes in stem and pocket structure[46]. Loss of hTR 5′ hairpin contacts to H/ACA proteins is compensated by an hTR-specific increase in protein binding to the 3′ hairpin[18]. The position of dyskerin defined by our structure explains why H/ACA RNP subunit exchange occurs in cell extracts for GAR1, NOP10 and NHP2, but not dyskerin[47]. The surprisingly limited protein–RNA interaction of the hTR 5′ hairpin implies that a wider breadth of RNA structures can assemble H/ACA proteins than has been predicted from the snoRNAs and scaRNAs, which conserve not only H/ACA protein binding but also requirements for dyskerin catalytic function[21].

The single subunit of TCAB1 in human telomerase contacts both the 3′ hairpin CAB box and dyskerin in the second set of H/ACA proteins (Fig. 5c, Extended Data Fig. 7g), giving a unifying explanation for its simultaneous identification as an RNA-independent dyskerin binding protein and a sequence-specific, direct RNA binding protein[48,49]. Unexpectedly, TCAB1 also contacts GAR1 (Fig. 5c, Extended Data Fig. 7g). Protein–protein interactions by TCAB1 explain why such a limited amount of CAB-box sequence can stabilize its association with H/ACA RNPs. Structure fitting of the archaeal H/ACA RNP, combined with sequence homology between the archaeal and human proteins, allowed mapping of human dyskeratosis congenita and Hoyeraal–Hreidarsson disease mutations in our structure[50] (Extended Data Fig. 8a–d). Notably, a large number of dyskerin mutations cluster at the dyskerin–dyskerin interface, which is also close to its RNA binding surface (Fig. 5d, Extended Data Fig. 7f, h). In addition, dyskerin R158 is near TCAB1, and NOP10 R34 and NHP2 V126 are near the P8 stem (Fig. 5c). Mutations at any of these newly evident interfaces would be expected to have more severe impact on hTR than snoRNAs or scaRNAs for three reasons: (a) crippled protein–RNA interaction of the hTR 5′ hairpin makes 5′ hairpin assembly more dependent on cross-hairpin dyskerin–dyskerin interactions; (b) hTR is uniquely reliant on enhanced H/ACA-protein affinity for the 3′-hairpin P8 stem-loop; and (c) hTR has one rather than two CAB box locations for TCAB1 association[18,46]. Our ability to map disease mutations to the H/ACA RNP structure expands the scope of telomerase deficiency mechanisms beyond decreased assembly of the core heterotrimer (dyskerin, NOP10, NHP2)[22] to include decreased dyskerin interaction with itself and with TCAB1 as well as with hTR.

Conclusions

The structure presented here provides, to our knowledge, the first architectural characterization of a telomerase holoenzyme from a multicellular species, addressing the longstanding question of vertebrate telomerase holoenzyme subunit content and arrangement. Our structure opens new opportunities for the design and screening of inhibitors targeting hTR folding and TERT interaction that will hold promise as general anti-cancer therapeutic agents[7].

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0062-x.

Received: 10 January 2018; Accepted: 28 March 2018;
Published online: 25 April 2018

1. Arnoult, N. & Karlseder, J. Complex interactions between the DNA-damage response and mammalian telomeres. *Nat. Struct. Mol. Biol.* **22**, 859–866 (2015).
2. Doksani, Y. & de Lange, T. The role of double-strand break repair pathways at functional and dysfunctional telomeres. *Cold Spring Harb. Perspect. Biol.* **6**, a016576 (2014).
3. Levy, M. Z., Allsopp, R. C., Futcher, A. B., Greider, C. W. & Harley, C. B. Telomere end-replication problem and cell aging. *J. Mol. Biol.* **225**, 951–960 (1992).
4. Holohan, B., Wright, W. E. & Shay, J. W. Cell biology of disease: telomeropathies: an emerging spectrum disorder. *J. Cell Biol.* **205**, 289–299 (2014).
5. Wegman-Ostrosky, T. & Savage, S. A. The genomics of inherited bone marrow failure: from mechanism to the clinic. *Br. J. Haematol.* **177**, 526–542 (2017).
6. Blackburn, E. H. & Collins, K. Telomerase: an RNP enzyme synthesizes DNA. *Cold Spring Harb. Perspect. Biol.* **3**, a003558 (2011).
7. Shay, J. W. Role of telomeres and telomerase in aging and cancer. *Cancer Discov.* **6**, 584–593 (2016).
8. Fu, D. & Collins, K. Distinct biogenesis pathways for human telomerase RNA and H/ACA small nucleolar RNAs. *Mol. Cell* **11**, 1361–1372 (2003).
9. Podlevsky, J. D. & Chen, J. J. Evolutionary perspectives of telomerase RNA structure and function. *RNA Biol.* **13**, 720–732 (2016).
10. Chan, H., Wang, Y. & Feigon, J. Progress in human and *Tetrahymena* telomerase structure determination. *Annu. Rev. Biophys.* **46**, 199–225 (2017).
11. Wu, R. A., Upton, H. E., Vogan, J. M. & Collins, K. Telomerase mechanism of telomere synthesis. *Annu. Rev. Biochem.* **86**, 439–460 (2017).
12. Gillis, A. J., Schuller, A. P. & Skordalakes, E. Structure of the *Tribolium castaneum* telomerase catalytic subunit TERT. *Nature* **455**, 633–637 (2008).
13. Mitchell, M., Gillis, A., Futahashi, M., Fujiwara, H. & Skordalakes, E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat. Struct. Mol. Biol.* **17**, 513–518 (2010).
14. Robart, A. R. & Collins, K. Human telomerase domain interactions capture DNA for TEN domain-dependent processive elongation. *Mol. Cell* **42**, 308–318 (2011).
15. Wu, R. A., Dagdas, Y. S., Yilmaz, S. T., Yildiz, A. & Collins, K. Single-molecule imaging of telomerase reverse transcriptase in human telomerase holoenzyme and minimal RNP complexes. *eLife* **4**, 10.7554 (2015).
16. Wenz, C. et al. Human telomerase contains two cooperating telomerase RNA molecules. *EMBO J.* **20**, 3526–3534 (2001).
17. Sauerwald, A. et al. Structure of active dimeric human telomerase. *Nat. Struct. Mol. Biol.* **20**, 454–460 (2013).
18. Egan, E. D. & Collins, K. Specificity and stoichiometry of subunit interactions in the human telomerase holoenzyme assembled *in vivo*. *Mol. Cell. Biol.* **30**, 2775–2786 (2010).
19. Jiang, J. et al. Structure of *Tetrahymena* telomerase reveals previously unknown subunits, functions, and interactions. *Science* **350**, aab4070 (2015).
20. MacNeil, D. E., Bensoussan, H. J. & Autexier, C. Telomerase regulation from beginning to the end. *Genes (Basel)* **7**, (64 (2016).
21. Hamma, T. & Ferré-D'Amaré, A. R. The box H/ACA ribonucleoprotein complex: interplay of RNA and protein structures in post-transcriptional RNA modification. *J. Biol. Chem.* **285**, 805–809 (2010).
22. Yu, Y.-T. & Meier, U. T. RNA-guided isomerization of uridine to pseudouridine–pseudouridylation. *RNA Biol.* **11**, 1483–1494 (2014).
23. Cohen, S. B. et al. Protein composition of catalytically active human telomerase from immortal cells. *Science* **315**, 1850–1853 (2007).
24. Li, L. & Ye, K. Crystal structure of an H/ACA box ribonucleoprotein particle. *Nature* **443**, 302–307 (2006).
25. Fu, D. & Collins, K. Purification of human telomerase complexes identifies factors involved in telomerase biogenesis and telomere length regulation. *Mol. Cell* **28**, 773–785 (2007).
26. Cristofari, G. et al. Human telomerase RNA accumulation in Cajal bodies facilitates telomerase recruitment to telomeres and telomere elongation. *Mol. Cell* **27**, 882–889 (2007).
27. Wu, R. A. & Collins, K. Human telomerase specialization for repeat synthesis by unique handling of primer-template duplex. *EMBO J.* **33**, 921–935 (2014).
28. Vogan, J. M. et al. Minimized human telomerase maintains telomeres and resolves endogenous roles of H/ACA proteins, TCAB1, and Cajal bodies. *eLife* **5**, 10.7554 (2016).
29. Bai, X.-C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. W. Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife* **4**, e11182 (2015).
30. Jiang, J. et al. The architecture of *Tetrahymena* telomerase holoenzyme. *Nature* **496**, 187–192 (2013).
31. Bajon, E., Laterreur, N. & Wellinger, R. J. A single templating RNA in yeast telomerase. *Cell Reports* **12**, 441–448 (2015).
32. Alves, D. et al. Single-molecule analysis of human telomerase monomer. *Nat. Chem. Biol.* **4**, 287–289 (2008).
33. Huang, J. et al. Structural basis for protein-RNA recognition in telomerase. *Nat. Struct. Mol. Biol.* **21**, 507–512 (2014).
34. Jacobs, S. A., Podell, E. R. & Cech, T. R. Crystal structure of the essential N-terminal domain of telomerase reverse transcriptase. *Nat. Struct. Mol. Biol.* **13**, 218–225 (2006).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

35. Petrova, O. A. et al. Structure and function of the N-terminal domain of the yeast telomerase reverse transcriptase. *Nucleic Acids Res.* **46**, 1525–1540 (2018).
36. Kim, N.-K. et al. Solution structure and dynamics of the wild-type pseudoknot of human telomerase RNA. *J. Mol. Biol.* **384**, 1249–1261 (2008).
37. Zhang, Q., Kim, N.-K. & Feigon, J. Architecture of human telomerase RNA. *Proc. Natl Acad. Sci. USA* **108**, 20325–20332 (2011).
38. Wang, Y., Yesselman, J. D., Zhang, Q., Kang, M. & Feigon, J. Structural conservation in the template/pseudoknot domain of vertebrate telomerase RNA from teleost fish to human. *Proc. Natl Acad. Sci. USA* **113**, E5125–E5134 (2016).
39. Theimer, C. A. et al. Structural and functional characterization of human telomerase RNA processing and cajal body localization signals. *Mol. Cell* **27**, 869–881 (2007).
40. Cash, D. D. & Feigon, J. Structure and folding of the *Tetrahymena* telomerase RNA pseudoknot. *Nucleic Acids Res.* **45**, 482–495 (2017).
41. Jansson, L. I. et al. Structural basis of template-boundary definition in *Tetrahymena* telomerase. *Nat. Struct. Mol. Biol.* **22**, 883–888 (2015).
42. Mitchell, J. R. & Collins, K. Human telomerase activation requires two independent interactions between telomerase RNA and telomerase reverse transcriptase. *Mol. Cell* **6**, 361–371 (2000).
43. Chen, J.-L., Opperman, K. K. & Greider, C. W. A critical stem-loop structure in the CR4-CR5 domain of mammalian telomerase RNA. *Nucleic Acids Res.* **30**, 592–597 (2002).
44. Bley, C. J. et al. RNA–protein binding interface in the telomerase ribonucleoprotein. *Proc. Natl Acad. Sci. USA* **108**, 20333–20338 (2011).
45. Chen, J.-L., Blasco, M. A. & Greider, C. W. Secondary structure of vertebrate telomerase RNA. *Cell* **100**, 503–514 (2000).
46. Egan, E. D. & Collins, K. An enhanced H/ACA RNP assembly mechanism for human telomerase RNA. *Mol. Cell. Biol.* **32**, 2428–2439 (2012).
47. Kittur, N., Darzacq, X., Roy, S., Singer, R. H. & Meier, U. T. Dynamic association and localization of human H/ACA RNP proteins. *RNA* **12**, 2057–2062 (2006).
48. Venteicher, A. S. et al. A human telomerase holoenzyme protein required for Cajal body localization and telomere synthesis. *Science* **323**, 644–648 (2009).
49. Tycowski, K. T., Shu, M. D., Kukoyi, A. & Steitz, J. A. A conserved WD40 protein binds the Cajal body localization signal of scaRNP particles. *Mol. Cell* **34**, 47–57 (2009).
50. Sarek, G., Marzec, P., Margalef, P. & Boulton, S. J. Molecular basis of telomere dysfunction in human genetic diseases. *Nat. Struct. Mol. Biol.* **22**, 867–874 (2015).

Acknowledgements We thank P. Grob, J. Fang, A. Chintangal, P. Tobias and H. Upton for technical support; X. Zhang and J. Vogan for establishing the suspension cell transfection protocol used in early stages of this work and sharing cell lines for in vivo studies and CAB-box mutant hTR; L. Kohlstaedt and QB3 mass spectrometry facility for analysis of purified telomerase; Y. Wang, J. Jiang and J. Feigon for sharing the coordinates of the *Tetrahymena* catalytic core and the modelled human t/PK; E. Rodina for sharing the *Hansenula polymorpha* TEN coordinates before publication; and H. Upton and A. Deshpande for comments on the manuscript. We thank the National Energy Research Scientific Computing Center supported by the Office of Science of the US Department of Energy for providing computing resources under contract number DE-AC02-05CH11231. This work was funded by N.I.H. grant GM054198 to K.C. T.H.D.N. is a Fellow of the University of California, Berkeley Miller Institute for Basic Research in Science. B.J.G. was supported by fellowships from the Swiss National Science Foundation (projects P300PA_160983, P300PA_174355). E.N. is a Howard Hughes Medical Investigator.

Reviewer information *Nature* thanks S. Scheres and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions K.C. and E.N. directed the study. T.H.D.N. developed the telomerase purification procedure and performed in vitro biochemical assays, negative-stain and cryo-EM specimen preparation, data collection, data processing and model fitting. J.T. performed in vivo experiments and assisted T.H.D.N with cell culture and biochemistry. R.A.W guided T.H.D.N with telomerase biochemistry and purification at the initial stage of the project. B.J.G. and D.T. helped T.H.D.N with data collection. T.H.D.N, K.C. and E.N. wrote the paper with contributions from all other authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0062-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0062-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to E.N. or K.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Telomerase reconstitution. Human telomerase was reconstituted in cells with the TERT expression plasmid pcDNA3.1-ZZ-Tev-TwinStrep-TERT expressing ZZ-SS-TERT, which has a Tobacco Etch Virus protease cleavage site between the tandem Protein A and TwinStrep tag modules, and the hTR expression plasmid pcDNA3.1-U3-hTR-HDV. HDV is the hepatitis delta virus ribozyme, which was previously shown to increase hTR accumulation[28,46]. Initially we used suspension HEK 293 cells lacking N-acetylglucosaminyltransferase I (GnTI)[51] and transfected the cells using polyethylenimine (PEI)[52]. These cells were cultured in FreeStyle 293 Expression medium (ThermoFisher). After 48 h of transfection, whole cell extract was prepared by CHAPS lysis[53]. We obtained populations of inactive and active telomerase RNP purified from these extracts. To enrich for the active telomerase RNP, we used adherent HEK 293T cells cultured in DMEM with GlutaMAX (ThermoFisher) supplemented with 10% FBS, 100 units per ml penicillin and 100 μg/ml streptomycin. The cells were transfected with TERT and hTR expression plasmids by calcium phosphate. After 48 h, the cells were harvested and resuspended in hypotonic lysis buffer (HLB) (20 mM HEPES NaOH pH 8, 2 mM $MgCl_2$, 0.2 mM EGTA, 10% glycerol, 0.1% NP40, 1 mM DTT and 0.1 mM PMSF). Whole-cell extract was prepared by three freeze–thaw cycles and adjusted to 300 mM before being cleared by centrifugation. The clarified extract was adjusted to 150 mM NaCl. For ΔTCAB1 telomerase and TERT–hTRmin, we used a previously characterized CAB-box mutant hTR (G414C)[25] and LhTRmin[28]. HEK 293 cell lines are long-term laboratory stocks and had the expected morphology and growth. They have not been authenticated or tested for mycoplasma recently.

Telomerase purification. The prepared whole-cell extract (from 120 plates of adherent cells or 6 l of cell suspension) was incubated with streptavidin agarose resin (Sigma-Aldrich) pre-bound to a 5′-biotinylated oligonucleotide[15] at room temperature for 3 h. The resin was washed with wash buffer (20 mM HEPES NaOH pH 8.0, 150 mM NaCl, 2 mM $MgCl_2$, 0.2 mM EGTA, 10% glycerol, 0.1% Igepal CA-630, 1 mM DTT) and eluted with a competitor oligonucleotide[15]. The eluate was subsequently incubated with MagStrep XT resin (IBA LifeSciences) for 2 h at 4 °C. The resin was washed with wash buffer and eluted in batch with biotin elution buffer (100 mM Tris HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 5–10 mM biotin, 2 mM $MgCl_2$, 0.1% Igepal CA-630, 10% glycerol, 1 mM DTT). We collected three sequential batch elution fractions for which the elution buffer was incubated with the resin for 30 min each time. For mass spectrometry, the eluate was combined and precipitated using trichloroacetic acid (TCA). The pellet was washed extensively with acetone and resuspended in water. The sample was analysed by MudPIT to identify the protein components present in the sample[54]. For cryo-EM studies, the final eluate was concentrated and buffer-exchanged with cryo-EM buffer (20 mM HEPES NaOH pH 8.0, 150 mM NaCl, 2 mM $MgCl_2$, 0.05% Igepal CA-630, 1% trehalose, 1 mM DTT). To prepare DNA-substrate bound telomerase, oligonucleotide $T_{14}AG_3$ (1 μM) was added in the MagStrep binding step, and excess DNA substrate was washed away. Substrate-bound telomerase was eluted as described above.

Immunoblotting. The purified telomerase samples (E1 and E2) from the CHAPS lysis protocol were resolved by bis-Tris SDS–PAGE. Proteins were transferred to a nitrocellulose membrane and blocked with 4% nonfat milk (Carnation) in TBS buffer (50 mM Tris pH 7.5, 150 mM NaCl) for 1 h at room-temperature. The membranes were incubated with primary antibodies overnight at 4 °C and washed with TBS before being probed with secondary antibodies (Invitrogen) in 4% nonfat milk in TBS for 1 h at room-temperature. We blotted for TERT, TCAB1, dyskerin and GAR1 on the same membrane, and NHP2 and NOP10 on a different membrane. The small sizes of NHP2 and NOP10 required a shorter protein transfer time than the other proteins. The membranes were washed extensively with TBS and visualized on a LI-COR Odyssey imager. The primary antibodies used were rabbit anti-dyskerin (Santa Cruz Biotechnology), mouse anti-TCAB1 (Sigma-Aldrich), rabbit anti-GAR1 (Novus), rabbit anti-NHP2 (Proteintech), rabbit anti-NOP10 (Abcam), mouse anti-Strep (Abcam), mouse anti-tubulin (Abcam) and rabbit IgG (Sigma-Aldrich).

Direct telomerase activity assays. Direct primer extension assays were performed using 1 μl sample in a 20-μl reaction volume containing 50 mM Tris-acetate pH 8.0, 4 mM $MgCl_2$, 5 mM DTT, 250 μM dTTP, 250 μM dATP, 5 μM unlabelled dGTP, 0.1 μM α-^{32}P dGTP (3,000 Ci/mmole, 10 mCi/ml, Perkin-Elmer), 500 nM 5′-$(T_2AG_3)_3$-3′ primer unless primer was pre-bound. Reactions were incubated at 30 °C for 40 min and stopped with TES buffer (50 mM Tris HCl pH 7.5, 20 mM EDTA, 0.2% SDS). The products were extracted and precipitated together with a 12-nt ^{32}P-end-labelled oligonucleotide as a recovery control and resolved on a denaturing polyacrylamide gel. The dried gels were exposed to a phosphorimager screen and imaged on a Typhoon Trio system (GE Healthcare).

HCT116 cell culture and genome editing. HCT116 cells were cultured in DMEM with GlutaMax (ThermoFisher) supplemented with 10% FBS and 100 μg/ml Primocin (Invivogen). Clonal cell lines with homozygous TERT disruption were generated previously[28]. Transgene integration at the *AAVS1* safe harbour locus was conducted as previously described[55] using CAGGS promoter for TERT expression and a neomycin resistance cassette for selection. Genome engineering in HCT116 cells was performed using the manufacturer's transfection protocol (Lipofectamine 3000, ThermoFisher) with media exchange at 24 h into selection with 200 μg/ml G418 and 100 μg/ml hygromycin. Cells were passaged continuously with selection medium for more than 3 months after selection for cells with the integrated TERT transgenes. HCT116 cells were obtained from Molecular and Cell Biology Department Tissue Culture Facility (UC Berkeley); this cell line was validated by short tandem repeat profiling with GenePrint and tested for mycoplasma by the Tissue Culture Facility.

Telomeric repeat amplification assays. Cell extract was prepared using the HLB method as described above and normalized to 2 mg/ml, which was quantified using the Bio-Rad Bradford protein assay. Dilutions were made of the whole cell extract, which were used for the PCR-based telomeric repeat amplification protocol (TRAP) as previously described[55]. Two microlitres with 400 ng total protein was used per 20 μl QTRAP reaction, consisting of 0.1 μg/μl TS primer (5′-AATCCGTCGAGCAGAGTT-3′), 0.1 μg/μl ACX primer (5′-GCGCGGCTTACCCTTACCCTTACCCTAACC-3′), and iTaq Universal Green Supermix (BioRad). Samples were incubated on a CFX96 Touch Real-Time PCR Detection System (Bio-Rad) at 30 °C for 30 min, followed by a hot start of 95 °C for 2 min. The samples were then cycled 35 times between 95 °C for 20 s and 61 °C for 90 s. Relative telomerase activity was calculated using ΔC_t normalized to the parental sample. Serial dilutions of 10, 30, and 100 ng were used for the hotTRAP protocol. The TS and ACX primers were used to amplify telomeric repeats labelled by incorporation of α-^{32}P-dGTP. TSNT (5′-AATCCGTCGAGCAGAGTTAAAAGGCCGAGAAGCGAT-3′) was used as the internal control (IC) for all samples, and NT (5′-ATCGCTTCTCGGCCTTTT-3′) was the dedicated TSNT reverse primer. Samples were resolved on a native gel. The gel was dried, exposed on a phosphorimager screen and imaged on a Typhoon scanner (GE Healthcare).

Southern blotting. Genomic DNA preparation was conducted as previously described[28]. Four to eight micrograms of DNA was digested for between 8 h and overnight with AluI and MboI at 37 °C before electrophoresis on a 0.7% agarose gel (Seakem ME agarose). The gel was dried and denatured in 1.5 M NaCl and 0.5 M NaOH for 1 h at 50 °C. The gel was then washed twice with 4× SSC + 0.1% SDS before prehybridization with Church's buffer (1% BSA, 1 mM EDTA, 0.5 M Na_3PO_4 pH 7.2, 7% SDS) for 30 min at 50 °C. ^{32}P-end-labelled probes were made of the telomeric repeat $(T_2AG_3)_3$ as well as the O'Generuler 1 kb plus ladder (ThermoFisher). These probes were both added and hybridized overnight at 50 °C. The gel was washed with 4× SSC + 0.1% SDS at 40 °C before exposing on a phosphorimager screen and imaging on a Typhoon scanner (GE Healthcare).

Negative-stain electron microscopy and data processing. For negative-stain grid preparation, 4 μl telomerase sample was incubated on glow-discharged 400-mesh copper grids, which had been previously coated with a homemade carbon film on nitrocellulose, for 90 s and stained with 2% w/v uranyl formate. Micrographs for the inactive and active fractions of telomerase were collected on an FEI Tecnai 12 transmission electron microscope operated at 120 kV and equipped with a TVIPS TemCam F416. Data was collected using the Leginon package[56] and data processing was done in RELION 1.4[57] unless otherwise indicated. Initial particle picking of ~1,500 particles was performed manually, particles were extracted with a 220^2 pixel box at a pixel size of 2.18 Å per pixel and subjected to reference-free 2D classification to obtain 2D initial references for auto-picking within RELION[58]. A total of 51,423 and 26,350 particles were picked for the inactive and active fractions, respectively. We performed two rounds of reference-free 2D classification for each dataset.

Micrographs for the enriched active fraction of intact holoenzyme, ΔTCAB1 telomerase, and TERT–hTRmin samples were collected on an FEI Tecnai F20 transmission electron microscope operated at 120 kV and equipped with a Gatan US4000 CCD camera. Data collection was done using the Leginon software package[56]. Initial particle picking for intact telomerase holoenzyme was done using DOG PICKER[59] within the APPION pipeline[60]. This yielded approximately 16,000 particles, which were extracted using a 400^2 pixel box at a pixel size of 1.5 Å per pixel and subjected to reference-free 2D classification to obtain initial 2D class averages. These 2D class averages were subsequently used as references for automatic picking within RELION[58] to yield 15,293 and 23,922 particles for intact telomerase holoenzyme and ΔTCAB1 mutant, respectively. Reference-free 2D classification was again performed on these particles. For the TERT–hTRmin complex, which is a much smaller RNP than telomerase holoenzyme, DOG PICKER was initially used to pick approximately 48,000 particles, which were extracted using a 180^2 pixel box at a pixel size of 1.5 Å per pixel. 2D class averages resulting from reference-free 2D classification of these particles were used as references for automatic picking in RELION to give 50,720 particles, which were subjected to another round of reference-free 2D classification. A comparison of the 2D class averages of intact telomerase holoenzyme, ΔTCAB1 mutant and TERT–hTRmin is shown in Extended Data Fig. 1e–h.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Cryo-electron microscopy. For cryo-EM analysis, 3 μl of the purified telomerase was applied to C-flat CF 4/2 grids (Protochips) that had been previously coated with a 6-nm-thick layer of homemade continuous carbon film and glow-discharged in amylamine[61] to overcome the preferred orientation of the telomerase holoenzyme on untreated carbon film. The grids were blotted for 4–5 s at 100% humidity and 4 °C and plunged into a mixture of liquid ethane and propane cooled by liquid nitrogen[62] using an FEI Vitrobot MK IV. The grids were loaded onto an FEI Titan Krios transmission electron microscope operated at 300 kV and equipped with a Gatan K2 Summit direct electron detector camera mounted behind a GIF Quantum energy filter. Images were collected automatically using the Serial-EM package[63] in super-resolution counting mode at a calibrated magnification of 43,477× (physical pixel size of 1.15 Å per pixel), with a total electron dose of 42 electrons per Å^2 during a total exposure time of 7 s, dose-fractionated into 28 movie frames. We used a slit width of 25 eV on the energy filter and a defocus range of 1.5–3.5 μm. A total of 11,654 micrographs were collected from two separate data collection sessions.

Image processing. Most steps of data processing were done in RELION 2.0[64] unless otherwise stated (Extended Data Fig. 3c). The movie stacks for each micrograph were corrected for drift and dose-weighted using MOTIONCOR2[65] running within Focus[66] during data collection. Contrast transfer function (CTF) parameters were estimated for the drift-corrected micrographs using GCTF[67]. Initial particle picking was performed on a subset of 350 micrographs using Gautomatch (from K. Zhang, MRC-LMB, Cambridge) and templates from a previous dataset collected on substrate-free telomerase at a different magnification (data not used and not shown in this study), resulting in approximately 32,000 particles. These particles were extracted with a 400^2 pixel box for reference-free 2D classification to obtain 2D references for automatic picking of all micrographs by Gautomatch. A total of 1,091,177 particles from 11,450 micrographs were picked, binned by 2 and extracted with a box size of 200^2 pixel. The data were initially processed in two batches, one for each data collection session. 3D classification into five classes was run for 30 iterations with an angular sampling of 7.5°, a regularization parameter T of 4 and a 60 Å low-pass filtered initial model obtained from the previous dataset of substrate-free telomerase (data not used and not shown in this study). The initial model used for the substrate-free telomerase was that of the published 30 Å negative stain reconstruction of human telomerase[17]. For dataset 1, a subset of 121,116 particles from one class with the best features was selected and un-binned. For dataset 2, we selected two 3D classes and subjected them to one round of reference-free 2D classification to remove more particle images showing contaminants or damaged telomerase complexes. A subset of 214,157 particles from this dataset was unbinned and combined with the subset of particles selected from dataset 1. The combined subset of 335,283 particles was refined and yielded a reconstruction that was hard to interpret due to the conformational heterogeneity between the two lobes and within each lobe. The angular assignments resulting from this refinement were used for focused 3D classification with signal subtraction for each lobe to resolve the conformational heterogeneity within each lobe[29]. The two lobes are henceforth referred to as the H/ACA lobe and the catalytic core, according to our subunit assignments using negative-stain EM (Extended Data Fig. 1e–h). The best class consisting of 103,363 particles for the catalytic lobe was refined to 7.8 Å resolution while that for the H/ACA lobe was refined to 8.8 Å resolution from 104,330 particles. These reconstructions show clear secondary structure features as expected for this resolution range. To further reduce conformational heterogeneity, we performed another round of alignment-free 3D classification, using the alignment parameters from the refinement run for each lobe, and refined the resulting best class. We obtained a 7.7 Å resolution reconstruction from 39,860 particles for the catalytic core and 8.2 Å resolution reconstruction from 36,210 particles for the H/ACA lobe. These reconstructions were used for model fitting (see below).

We also performed alignment-free 2D classification on the original (unsubtracted) particles using the orientations determined by the 3D refinement of each lobe to assess the relative orientation of the other lobe (Extended Data Fig. 4a). For both cases, the other lobe adopts a wide range of conformations relative to the lobe that is aligned, clearly showing that the quality of any reconstruction for the whole structure is limited by both the continuous movements between the two lobes relative to each other and the conformational heterogeneity within each lobe. To obtain a better reconstruction encompassing both lobes, we performed global 3D classification for the subset of 335,283 particles before signal subtraction (25 iterations, $T = 8$, angular sampling of 7.5°) (Extended Data Fig. 3c). As expected for a particle population exhibiting continuous motion, it resulted in four classes with nearly equal distribution and the two lobes adopting different conformations relative to each other. We refined each of these classes and obtained 10.2 Å resolution for the best class.

All refinements were performed using fully independent data half-sets (gold-standard refinement) and resolutions are reported based on the Fourier shell correlation (FSC) = 0.143 criterion[68,69] (Extended Data Fig. 4b). FSC curves were calculated with a soft mask. For visualization, maps were corrected for the modulation transfer function of the detector and sharpened by applying a negative B factor as determined by the post-processing function of RELION. Local resolution (Extended Data Fig. 4d, e) was calculated within RELION 2.0. For model versus map FSC curves calculations, we converted the catalytic core and H/ACA RNP models into density using EMAN[70]. Owing to small regions of unmodelled density in our maps (Extended Data Fig. 4f), we prepared a soft mask (6-pixel soft edge) from the model and extracted the cryo-EM maps of each model using this mask. FSC model versus map was calculated using Xmipp[71] (Extended Data Fig. 4c).

Model fitting. Initial fitting of all X-ray structures, NMR ensembles and homology models was performed manually by visual inspection of the best maps of the individual two lobes, followed by fit optimization in Chimera[72] (Extended Data Table 2b, Supplementary Data 1 and Supplementary Video 1). The fitted structure for each lobe was subsequently rigid-body fitted into the overall map resulting from refinement of the best 3D class from global 3D classification. A homology model for TCAB1 WD40 domain was obtained using Phyre[73] based on the structure of coatomer subunit alpha in the COPI coat linkage structure[74] (PDB 5A1V) and fitted into the density located by comparisons of the 2D class averages for wild-type and ΔTCAB1 telomerase (Extended Data Fig. 1e–h). The *Tetrahymena* TEN domain (PDB 2B2A) and the TRBD, RT and CTE domains of the *Tribolium* TERT[13] (PDB 3KYL) were fitted by rigid-body fitting into the map separately. We compared the fittings of the *Tribolium* TRBD and the medaka TRBD (PDB 4O26) (Extended Data Fig. 7e). The former has a truncated TRBD, while the latter is closely related to the human TRBD in sequence. Some of the medaka TRBD extensions close to the RNA binding surface fit our map, while those in the periphery of the domain (including residues 403–425 and residues 545–572) did not. We removed extensions that did not fit from the medaka TRBD and replace the *Tribolium* TRBD with this truncated version. Our substrate can form a 6-bp duplex with hTR template region. Thus we truncated the DNA/RNA duplex model in the duplex-bound *Tribolium* TERT (PDB 3KYL) to 6 bp. Overall, the relative orientations of the domains and the DNA–RNA duplex are very similar to that of the crystal structure. For the H/ACA RNP, the RNA used in the single-hairpin archaeal crystal structure (PDB 2HVY) was truncated to a 9-bp hairpin flanked by the 3′ ACA motif. We used this model to fit both H/ACA copies because no structure of H/ACA proteins bound to the H box is available. Fitted structures of available hTR domains include the human P2b–P3 pseudoknot[36] (PDB 2K95), P2a–J2a/b–P2b[37] (PDB 2L3E), P2a.1[38], P8 stem (or CR7)[39] (PDB 2QH2), and the medaka TRBD-bound CR4/5[33] (PDB 4O26). The P4 and P7 stems bound by the H/ACA proteins were adapted from the truncated archaeal hairpin RNA model, as mentioned above. We used COOT[75] to place idealized A-form RNA helices in the map regions assigned to the P1, P4.2, P4.2 and P6b stems.

Map and model visualization. Maps were visualized in Chimera[72] and all model illustrations were prepared using either Pymol (https://www.pymol.org) or Chimera.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

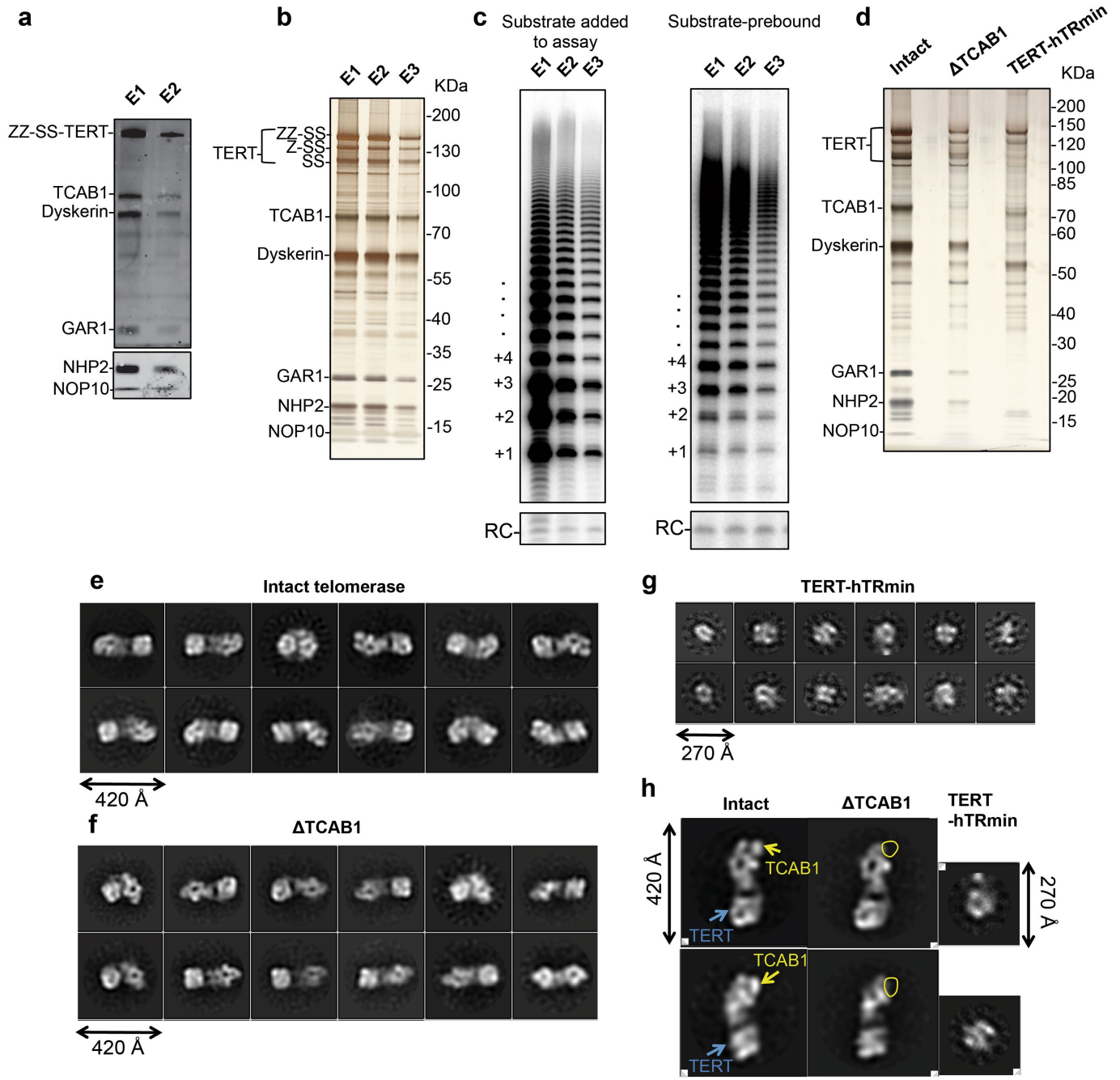
Data availability. Figure source data are provided in Supplementary Fig. 1. Cryo-EM maps of the catalytic core, H/ACA lobe and overall human telomerase holoenzyme have been deposited with the Electron Microscopy Data Bank under accession numbers EMD-7518, EMD-7519 and EMD-7521, respectively. The PyMol session used for fitting of homology models into the EM density is provided as Supplementary Information.

51. Reeves, P. J., Callewaert, N., Contreras, R. & Khorana, H. G. Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proc. Natl Acad. Sci. USA* **99**, 13419–13424 (2002).
52. Durocher, Y., Perret, S. & Kamen, A. High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. *Nucleic Acids Res.* **30**, E9 (2002).
53. Kim, N. W. et al. Specific association of human telomerase activity with immortal cells and cancer. *Science* **266**, 2011–2015 (1994).
54. Schirmer, E. C., Yates, J. R. III & Gerace, L. MudPIT: A powerful proteomics tool for discovery. *Discov. Med.* **3**, 38–39 (2003).
55. Sexton, A. N. et al. Genetic and molecular identification of three human TPP1 functions in telomerase action: recruitment, activation, and homeostasis set point regulation. *Genes Dev.* **28**, 1885–1899 (2014).
56. Suloway, C. et al. Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151**, 41–60 (2005).
57. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
58. Scheres, S. H. W. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* **189**, 114–122 (2015).
59. Voss, N. R., Yoshioka, C. K., Radermacher, M., Potter, C. S. & Carragher, B. DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. *J. Struct. Biol.* **166**, 205–213 (2009).
60. Lander, G. C. et al. Appion: an integrated, database-driven pipeline to facilitate EM image processing. *J. Struct. Biol.* **166**, 95–102 (2009).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

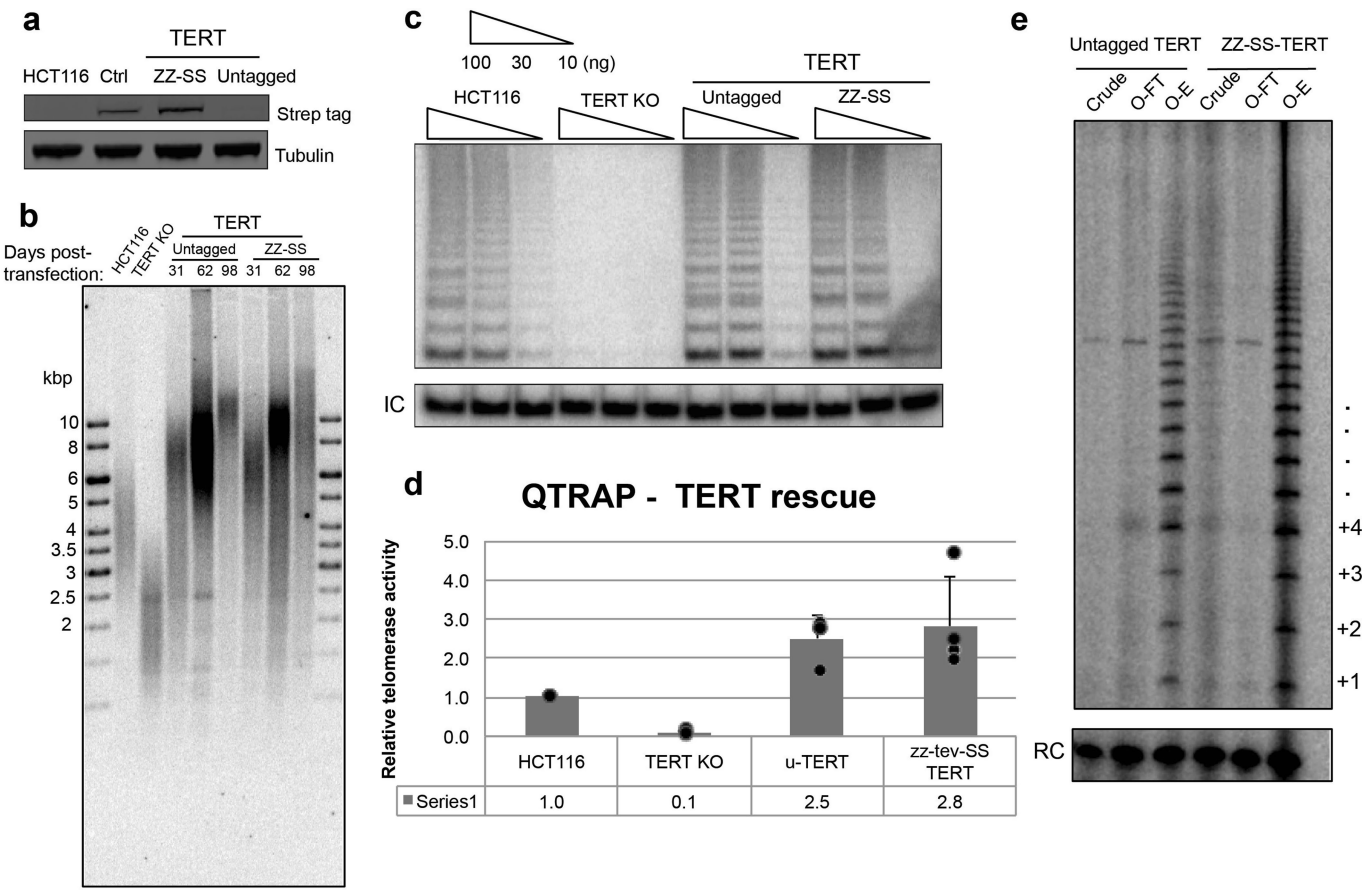
61. Nguyen, T. H. D. et al. The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* **523**, 47–52 (2015).
62. Tivol, W. F., Briegel, A. & Jensen, G. J. An improved cryogen for plunge freezing. *Microsc. Microanal.* **14**, 375–379 (2008).
63. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
64. Kimanius, D., Forsberg, B. O., Scheres, S. H. W. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, 18722 (2016).
65. Zheng, S. Q., Palovcak, E., Armache, J.-P., Cheng, Y. & Agard, D. A. MotionCor2: anisotropic correction of beam-induced motion for improved single particle electron cryo-microscopy. *Nat. Methods* **14**, 331–332 (2017).
66. Biyani, N. et al. Focus: The interface between data collection and data processing in cryo-EM. *J. Struct. Biol.* **198**, 124–133 (2017).
67. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
68. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
69. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* **9**, 853–854 (2012).
70. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
71. Scheres, S. H. W., Núñez-Ramírez, R., Sorzano, C. O. S., Carazo, J. M. & Marabini, R. Image processing for electron microscopy single-particle analysis using XMIPP. *Nat. Protocols* **3**, 977–990 (2008).
72. Goddard, T. D., Huang, C. C. & Ferrin, T. E. Visualizing density maps with UCSF Chimera. *J. Struct. Biol.* **157**, 281–287 (2007).
73. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols* **10**, 845–858 (2015).
74. Dodonova, S. O. et al. Vesicular transport. A structure of the COPI coat and the role of coat proteins in membrane vesicle assembly. *Science* **349**, 195–198 (2015).
75. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
76. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



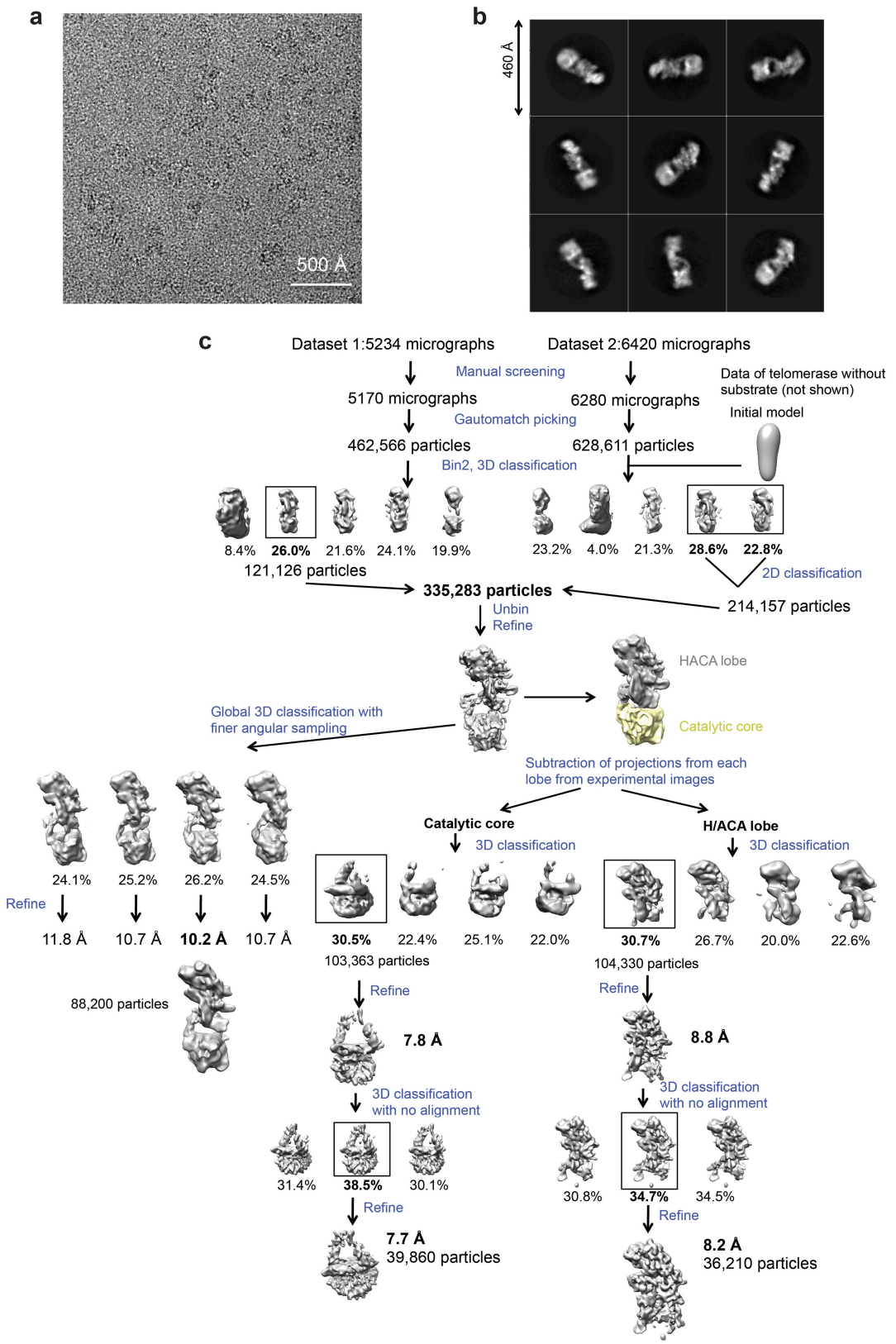
Extended Data Fig. 1 | Protein identification by immunoblotting, enriching active telomerase, substrate pre-binding, and comparison of intact, ΔTCAB1, and TERT–hTRmin RNPs. a, Immunoblotting of TERT, TCAB1, dyskerin, GAR1, NHP2 and NOP10 in telomerase purified after CHAPS lysis protocol as shown in Fig. 1b. We used primary antibodies against each protein, except ZZ-SS-TERT, for which we used rabbit IgG. Owing to the wide range of molecular weights of the proteins in our sample, TERT, TCAB1, dyskerin and GAR1 were detected in one blot, while NHP2 and NOP10 were detected in a separate blot. The use of the same sample to probe all proteins was performed only once, but TERT, dyskerin and TCAB1 were also probed individually twice. **b**, Silver-stained SDS–PAGE gel of purified telomerase fractions obtained from adherent cells lysed using the hypotonic lysis method, which enriches active telomerase. This experiment was repeated more than five times with similar results. **c**, Direct primer-extension assays of the purified telomerase fractions shown in **b**, confirming that E1 is no longer inactive (left), and of the substrate-bound purified telomerase fractions with additional DNA substrate omitted from the assays (right). The activity observed confirmed that purified telomerase contains the DNA substrate. The activity assays with substrate added were repeated over five times and the activity assays with substrate pre-bound were repeated twice. All repeats showed similar results. **d**, Silver-stained SDS–PAGE gel of purified intact and ΔTCAB1 telomerase and TERT–hTRmin telomerase prepared for subunit assignments. This experiment was done only once to provide a direct comparison between these different purified telomerase complexes. **e**–**g**, Negative-stained 2D class averages of intact and ΔTCAB1 telomerase and TERT–hTRmin, respectively. **h**, Comparison of representative 2D class averages of intact and ΔTCAB1 telomerase and TERT–hTRmin showing the inferred localization of TCAB1 and TERT. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



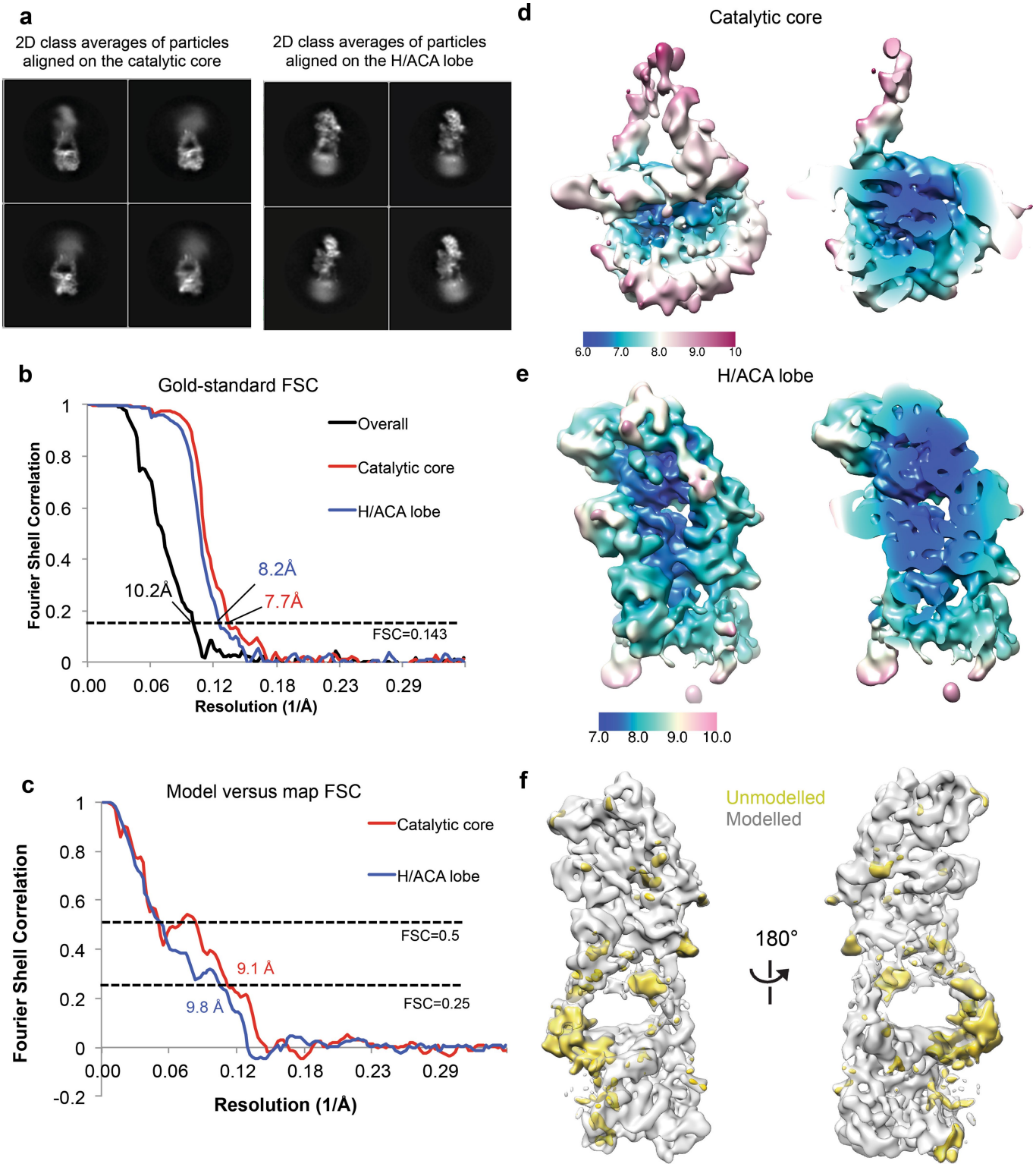
Extended Data Fig. 2 | Cellular function of the tagged TERT used for structural analysis. a, Western blot detection of ZZ-SS-TERT in TERT knockout (KO) cells rescued by untagged TERT or ZZ-SS-TERT expression. Whole-cell extracts were probed using Strep antibody. HCT116 is the parental cell line. Lysate prepared from HEK 293T cells transiently transfected with ZZ-SS-TERT and hTR was used as a positive control (Ctrl). Tubulin was detected as a loading control. This experiment was performed only once to confirm the success of ZZ-SS-TERT incorporation into the HCT116 TERT KO cells. **b**, Telomeric restriction fragment analysis of HCT116 parental cells, TERT KO cells (before senescence), and TERT KO cells rescued with untagged or ZZ-SS-TERT transgene. Transgene-expressing cells were sampled at 31, 62 and 98 days post-transfection with transgene vectors. This experiment was performed twice with similar results. **c**, TRAP assay detection of telomerase activity in HCT116 parental cells, TERT KO cells, and TERT KO cells rescued with untagged TERT or ZZ-SS-TERT transgene. Whole-cell extracts were normalized by total protein concentration and assayed at 100, 30 or 10 ng of total protein per reaction. IC, internal control. This experiment was repeated four times with similar results. **d**, Quantification of Q-TRAP assay detection of telomerase activity in HCT parental cells, TERT KO cells and TERT KO cells rescued with untagged TERT or ZZ-SS-TERT transgene. Error bars were calculated by taking the s.d. of the average ΔCt from four time points. Data points were shown as overlays. **e**, Direct primer-extension assay of telomerase after template-complementary oligonucleotide purification from extracts of TERT KO cells rescued by untagged TERT or ZZ-SS-TERT transgene. Assays were performed on clarified cell lysate (crude), flow-through (O-FT) and elution (OE) using equivalent amounts of cell extract. This experiment was performed only once to re-confirm the results in **c**. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



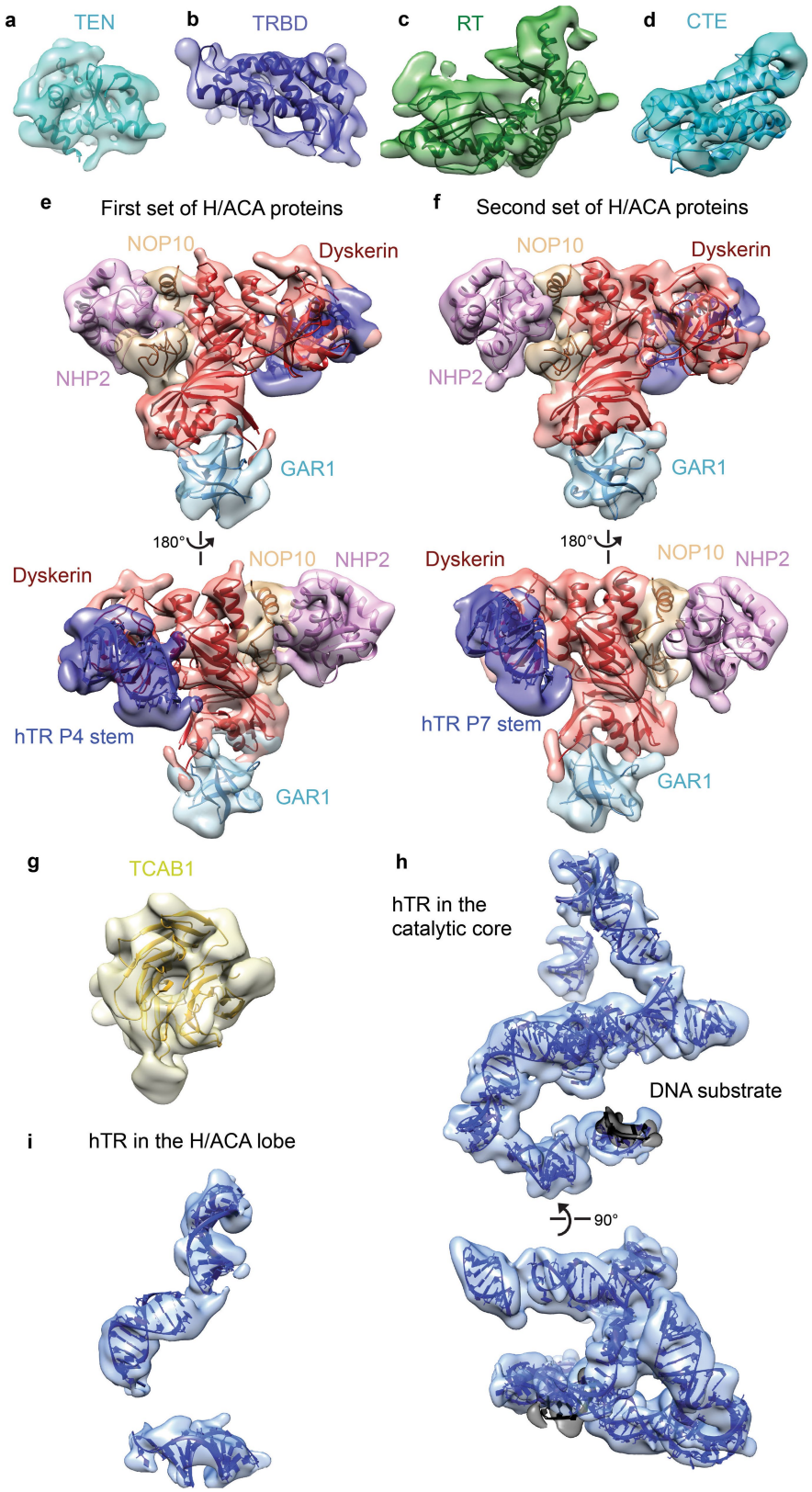
Extended Data Fig. 3 | Image processing procedures. a, Representative raw micrograph. We collected a total of 11,654 micrographs for this study. **b**, Representative 2D class averages obtained from reference-free 2D classification. **c**, Data processing strategy used in this study.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



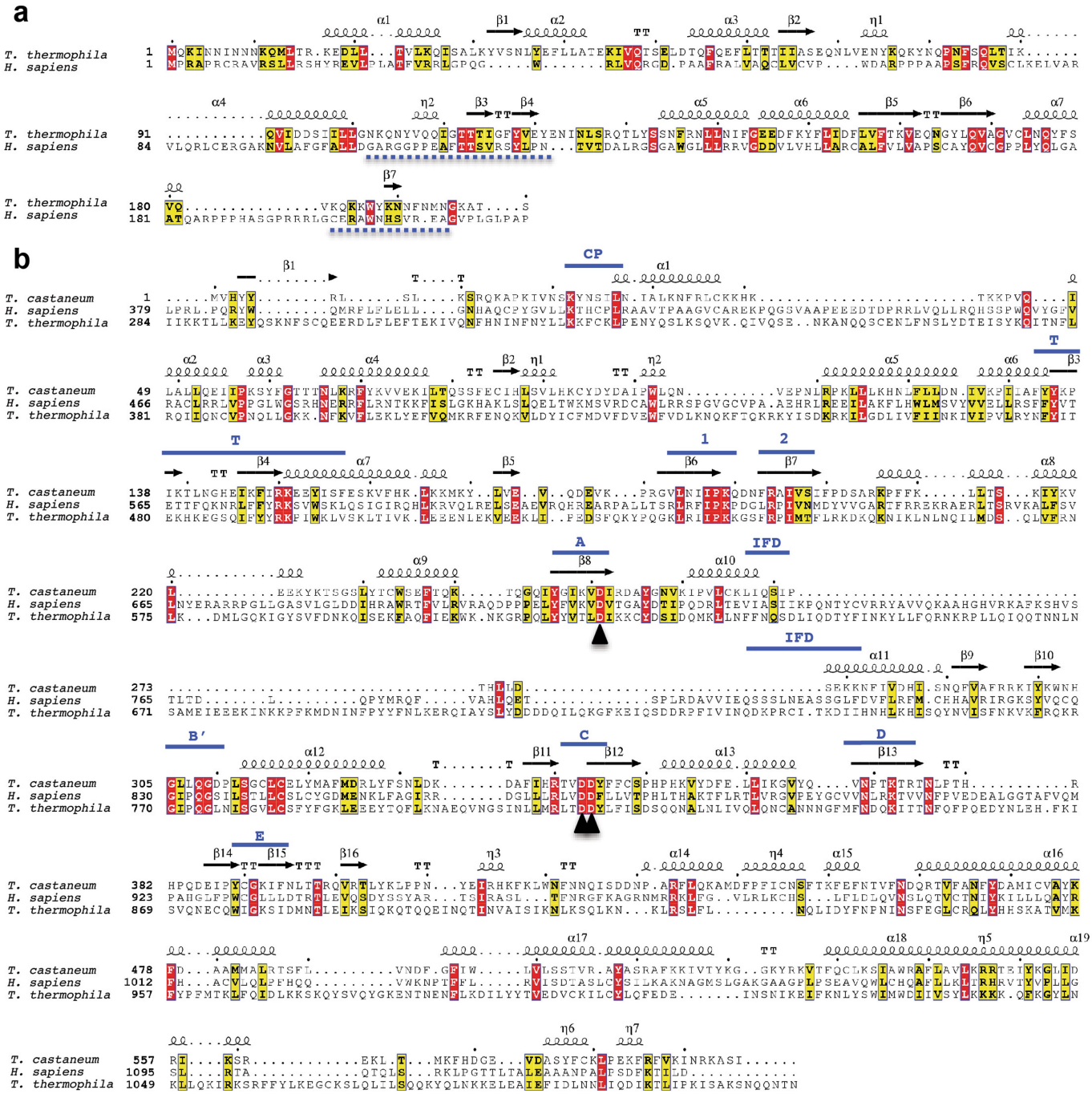
Extended Data Fig. 4 | Resolution estimation and analysis of the flexibility of the complex. a, Representative 2D class averages obtained from 2D classification without alignment of particles that were aligned on either the catalytic core or the H/ACA lobe. For both cases, the other lobe adopts a wide range of conformations, as illustrated by the blurriness of the density. **b**, FSC curves for the overall map and the maps of the catalytic core and H/ACA lobe resulting from focused classification with signal subtraction and gold-standard refinement. **c**, Model versus map FSC curves for the catalytic core and the H/ACA RNP. We fitted only homology models as rigid bodies into the map and did not perform model coordinate refinement owing to the limited resolutions of the maps. Therefore, we used a lower FSC threshold of 0.25 for resolution estimates. **d**, **e**, Local resolution for the catalytic lobe (**d**) and the H/ACA lobe (**e**) estimated by RELION 2.0[64]. Most of the catalytic core is resolved at 6–8 Å while most of the H/ACA lobe is resolved at 7–9 Å. **f**, Front (left) and back (right) views of the reconstruction showing modelled (grey) and unmodelled (gold) density. Most of the unmodelled density corresponds to single-stranded RNA regions or RNA bulges, and human protein extensions that cannot be built de novo at this resolution.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



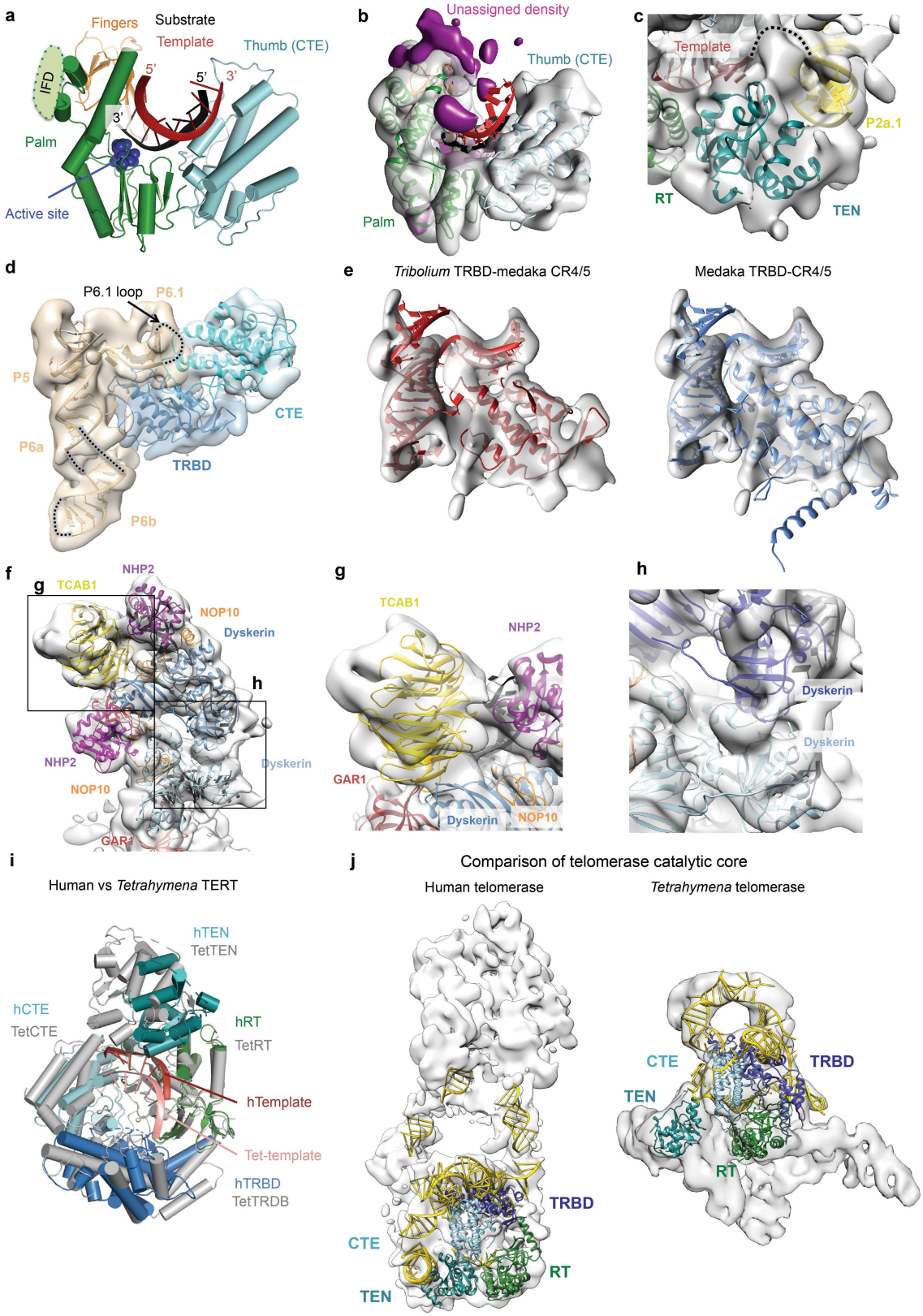
Extended Data Fig. 5 | Fittings of proteins and RNA into the cryo-EM map. a–d, Domains of TERT. **a**, The TEN domain from *Tetrahymena*[34] (PDB 2B2A). **b**, The truncated medaka TRBD domain[33] (PDB 4O26). **c**, **d**, The RT and CTE domains from *Tribolium*[13] (PDB 3KYL). **e**, Front (top) and back (bottom) views of the 5′ hairpin set of H/ACA proteins (dyskerin, red; GAR1, cyan; NOP10, wheat; NHP2, pink) bound to P4 stem (dark blue) fit by the archaeal H/ACA RNP[24] (PDB 2HVY). **f**, Front (top) and back (bottom) views of the 3′ hairpin set of H/ACA proteins using the same model and colour scheme as **e**. **g**, Homology model of TCAB1 WD40 domain. **h**, Front (top) and bottom (bottom) views of hTR in the catalytic core. **i**, hTR in the H/ACA lobe.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Sequence alignment of TERT with secondary structure assignments based on known structures. a, Sequence alignment of *Tetrahymena* and human TEN domains. The secondary structure assignments of the *Tetrahymena* TEN domain[34] (PDB 2B2A) are shown above the aligned sequences. Regions removed before fitting are indicated with dashed lines below the sequences. **b**, Sequence alignment of the *Tribolium*, human and *Tetrahymena* TERT, with the latter two N-terminally truncated to match *Tribolium*. Secondary structure assignments of the *Tribolium* TERT are shown on top, with conserved motifs labelled in blue. Throughout the figure, the η symbol refers to a 3_{10}-helix. Strict β-turns and strict α-turns are displayed as TT and TTT. The three catalytic aspartic acids are indicated with black arrowheads. ESpript was used to generate this figure[76].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

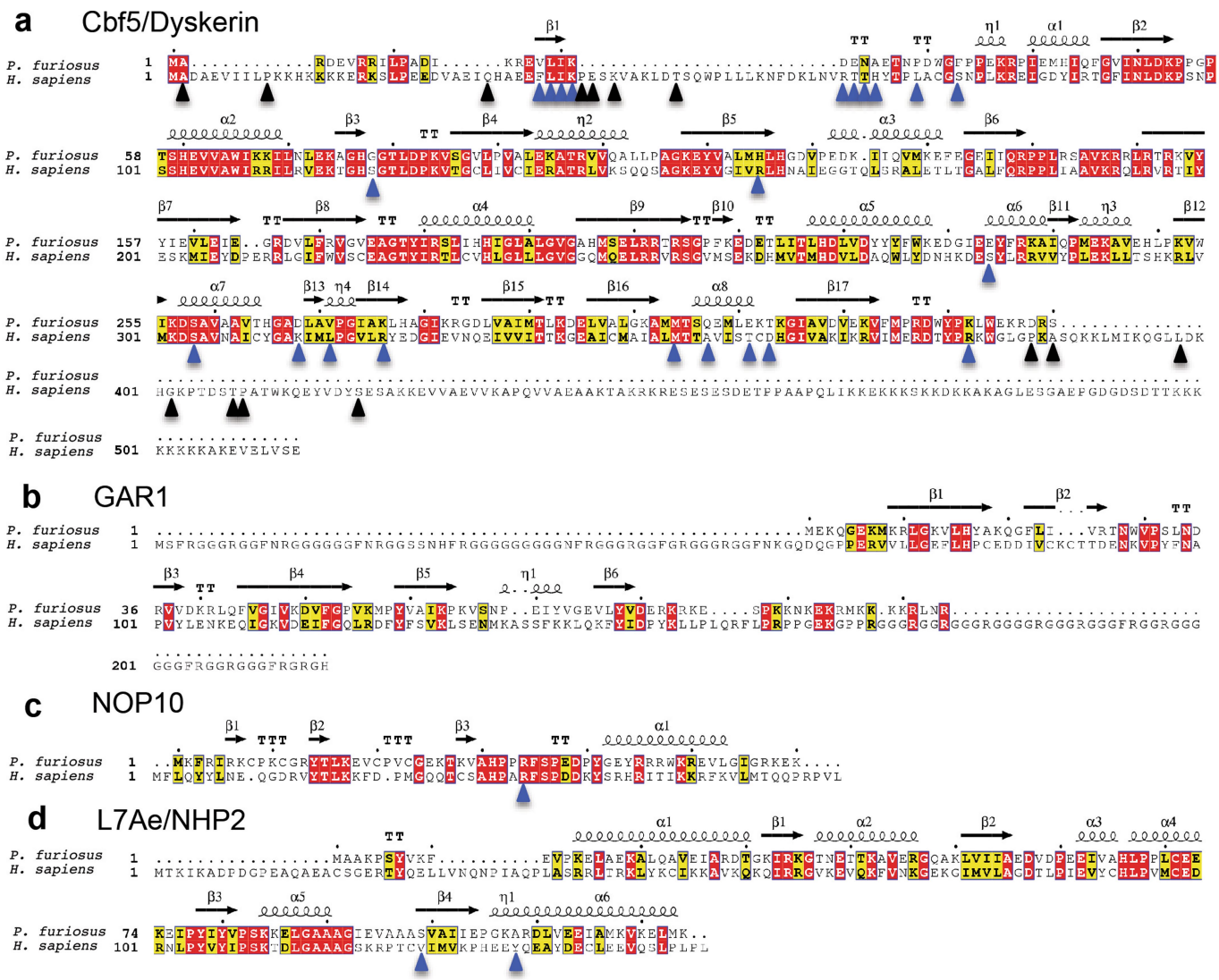


Extended Data Fig. 7 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 7 | Selected protein–protein and protein–RNA interactions in telomerase holoenzyme and comparisons between human and *Tetrahymena* TERT. a, Interactions between the RT and CTE domains of TERT and the substrate–template duplex. The RT domain is divided into two subdomains, the palm (green) and fingers (orange), that are commonly observed in retroviral reverse transcriptases. The CTE (cyan) is the putative thumb. The IFD insertion that is missing in the *Tribolium* TERT is indicated. **b**, Region of the cryo-EM reconstruction shown in **a**. Unassigned density close to the IFD insertion is highlighted in magenta. **c**, Cryo-EM density of the TEN domain in the same view as that in Fig. 4b. Connecting density is observed between the template region and the P2a.1 stem. **d**, Map of the CR4/5 three-way junction (wheat) and the nearby TERT domains highlighting the position of the P6.1 loop near the interface of the CTE (cyan) and TRBD (blue) domains of TERT. This loop was not ordered in medaka CR4/5 bound to the TRBD alone[33]. **e**, Comparison of the *Tribolium* (left) and medaka (right) TRBD with the medaka CR4/5 domain of hTR[13,33]. Extensions of the medaka TRBD that did not fit the map were truncated for visualization. **f**, Cryo-EM map with H/ACA components fitted. **g**, **h**, Detailed views of regions boxed in **f** show TCAB1 interactions with dyskerin, GAR1 and the P8 stem-loop (**g**), and interactions between the two dyskerin molecules (**h**), where a cluster of DC mutations are found (Fig. 5d). **i**, Comparison of the human and *Tetrahymena* TERT superposed on the RT domain. Domains of human TERT are coloured as in Fig. 1a, while *Tetrahymena* TERT is coloured grey. The bound human and *Tetrahymena* templates are coloured dark and light red, respectively. **j**, Comparison of human and *Tetrahymena*[19] catalytic cores fitted into the corresponding cryo-EM maps. Domains of TERT were coloured as in Fig. 1a and TER is coloured yellow. We used the catalytic core and H/ACA lobe densities resulting from our focused classification/refinement for the human telomerase and the overall 9.4 Å *Tetrahymena* telomerase map (EMD-6442).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 8 | Sequence alignments of H/ACA proteins with secondary structure assignments based on known structures. **a**–**d**, Sequence alignments of *Pyrococcus furiosus* (archaeal) and human Cbf5/dyskerin (**a**), GAR1 (**b**), NOP10 (**c**), and L7Ae/NHP2 (**d**). Secondary structure assignments displayed on the top are from the archaeal H/ACA RNP structure[24] (PDB 2HVY). The η symbol refers to a 3_{10}-helix. Strict β-turns and strict α-turns are displayed as TT and TTT, respectively. Known human dyskeratosis congenita and Hoyeraal–Hreidarsson disease mutations[50] in H/ACA proteins are indicated with arrowheads. Blue arrowheads indicate residues that can be mapped onto the archaeal structure and black arrowheads indicate residues that were not mapped. ESpript was used to generate this figure[76].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Mass spectrometry analysis of the purified telomerase sample

Protein name	Sequence count	Sequence coverage (%)	Length (aa)	MW (KDa)
NHP2 (H/ACA subunit 2)	2	19	153	17.2
NOP10 (H/ACA subunit 3)	1	17.2	64	7.7
Nuclease-sensitive element binding protein	4	13.6	324	35.9
Methylosome subunit pICln	2	13.5	237	26.2
TERT	11	10.6	1132	127
Insulin-like growth factor 2	3	8.6	548	60.1
TCAB1/WDR79/WRAP53beta	3	8.6	548	59.3
squamous cell carcinoma antigen recognised by T-cells 3 (SART3)	6	8	963	110
Importin subunit α-1	3	7.8	529	57.9
Dyskerin (H/ACA subunit 4)	5	7.2	514	57.7
hnRNP A2/B1	2	7.7	313	33.9
hnRNP A3	1	5.8	378	39.6
Protein arginine N-methyltransferase 5	1	5.6	531	61.1
ATP-dependent RNA helicase DHX36	2	2.7	1008	114.8

Proteins are ranked according to their percentage of sequence coverage eliminating keratins, histones, and abundant cytoskeletal proteins. Only proteins with sequence coverages of above 5% or known telomerase-associated proteins (red background) with any level of detection are shown.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Summary of cryo-EM data collection and modelling of protein and RNA subunits

a. Cryo-EM data collection

	Catalytic Core (EMDB-7518)	H/ACA lobe (EMDB-7519)	Overall (EMDB-7521)
Data collection and processing			
Magnification	300	300	300
Voltage (kV)			
Electron exposure (e–/Å^2)	42	42	42
Defocus range (μm)	1.5-3.5	1.5-3.5	1.5-3.5
Pixel size (Å)	1.15	1.15	1.15
Symmetry imposed	C1	C1	C1
Initial particle images (no.)	1,091,177	1,091,177	1,091,177
Final particle images (no.)	39,860	36,210	88,200
Map resolution (Å)	7.7	8.2	10.2
FSC threshold	0.143	0.143	0.143
Map resolution range (Å)	6-9	7-9	

b. Modelling of protein and RNA subunits

Protein/RNA	Total residues	M.W. (KDa)	Domain	PDB code or modelling	Reference
TERT	1132	127	TEN	2B2A (*Tetrahymena* model)	34
			TRBD	4O26 (*medaka* model)	33
			RT	3KYL (Tribolium model)	13
			CTE		
TCAB1	548	59	β-propeller domain	homology model by Phyre	
Dyskerin	514	57		2HVY (archaeal model)	24
GAR1	217	21			
NOP10	64	8			
NHP2	153	17			
hTR	451	138	P1	A-form double helix	
			P2a.1	model from Chan et al, 2017	10, 38
			P2a-J2a/b-P2b	2L3E	37
			template	3KYL	13
			P3	2K95	36
			P4	2HVY as part of the H/ACA RNP	24
			P4.1	A-form double helix	
			P4.2	A-form double helix	
			P5-P6.1-P6a	4O26 (medaka model)	33
			P6b	A-form double helix	
			P7	2HVY as part of the H/ACA RNP	24
			P8	2QH2	39

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0083-5

Structural basis for dual-mode inhibition of the ABC transporter MsbA

Hoangdung Ho[1], Anh Miu[2,9], Mary Kate Alexander[3,9], Natalie K. Garcia[4,9], Angela Oh[1], Inna Zilberleyb[5], Mike Reichelt[6], Cary D. Austin[6], Christine Tam[5], Stephanie Shriver[5], Huiyong Hu[7], Sharada S. Labadie[7], Jun Liang[7], Lan Wang[7], Jian Wang[8], Yan Lu[8], Hans E. Purkey[7], John Quinn[2], Yvonne Franke[5], Kevin Clark[2], Maureen H. Beresini[2], Man-Wah Tan[3], Benjamin D. Sellers[7], Till Maurer[1], Michael F. T. Koehler[7], Aaron T. Wecksler[4], James R. Kiefer[1], Vishal Verma[7], Yiming Xu[2], Mireille Nishiyama[3], Jian Payandeh[1,3,10]* & Christopher M. Koth[1,3,10]*

The movement of core-lipopolysaccharide across the inner membrane of Gram-negative bacteria is catalysed by an essential ATP-binding cassette transporter, MsbA. Recent structures of MsbA and related transporters have provided insights into the molecular basis of active lipid transport; however, structural information about their pharmacological modulation remains limited. Here we report the 2.9 Å resolution structure of MsbA in complex with G907, a selective small-molecule antagonist with bactericidal activity, revealing an unprecedented mechanism of ABC transporter inhibition. G907 traps MsbA in an inward-facing, lipopolysaccharide-bound conformation by wedging into an architecturally conserved transmembrane pocket. A second allosteric mechanism of antagonism occurs through structural and functional uncoupling of the nucleotide-binding domains. This study establishes a framework for the selective modulation of ABC transporters and provides rational avenues for the design of new antibiotics and other therapeutics targeting this protein family.

An obstacle in addressing the need for new antibiotics targeting Gram-negative bacteria is the requirement that these molecules penetrate the outer membrane[1,2]. A modular glycolipid known as lipopolysaccharide (LPS) resides exclusively within the outer membrane and provides a highly impermeable barrier[3]. A specialized transport system moves LPS to the cell surface, but LPS biogenesis occurs in the cytoplasm[4–6]. There, oligosaccharides are ligated to a lipid A membrane anchor, forming core-LPS[4]. Because LPS cannot spontaneously exchange across the inner membrane at a physiologically relevant rate, MsbA, an essential ATP-binding cassette (ABC) transporter, must carry out its vectorial transport[6,7]. MsbA is a homodimeric member of the B-family of ABC transporters, in which ATP hydrolysis by the intracellular nucleotide-binding domains (NBDs) powers unidirectional LPS transport[8]. A recent framework provided by cryo-electron microscopy (cryo-EM) supports the existence of inward-facing and outward-facing conformations of MsbA associated with a 'trap and flip' transport mechanism[9]. Notably, loss of MsbA causes rapid membrane perturbation and cell death[7], identifying it as a potential antibacterial target.

The discovery of potent and selective modulators of ABC transporters remains challenging, limiting our mechanistic understanding of their pharmacological regulation. This is exemplified by MsbA, a largely unexplored antibacterial target that offers a potential new strategy to overcome existing antibiotic-resistance mechanisms. We set out to identify selective MsbA antagonists and found a quinoline class of molecules with bactericidal activity. Structural studies revealed that these inhibitors target a previously unknown allosteric transmembrane site to trap MsbA in an inward-facing, LPS-bound state that prevents transition to the outward-facing conformation. Upon inhibitor binding, the NBDs also become uncoupled and adopt an asymmetric conformation not previously observed in other ABC transporter structures. Overall, this study presents a structural template for selective small molecule inhibition within the ABC transporter superfamily.

Discovery of selective MsbA inhibitors

G592 is a quinoline compound with potent activity on purified *Escherichia coli* MsbA (*Ec*MsbA) that was identified from a library of approximately 3,000,000 small molecules and specifically inhibits the activity of MsbA in cells (Fig. 1a, b, Supplementary Information Section 1 and Extended Data Figs. 1, 2). Inhibition was enantiomer-selective as the (*R*)-(−) form, G593, displayed almost tenfold lower activity than G592. Efforts aimed at improving the potency of G592 led to the discovery of G247 and G907 (Fig. 1a, b; G247 half-maximum inhibitory concentration (IC_{50}) = 5 nM, G907 IC_{50} = 18 nM). G247 and G907 treatment mirrored the distinctive membrane elaboration phenotype of *E. coli* cells depleted of MsbA[7] (Fig. 1c and Extended Data Fig. 2) and inhibited the growth of wild-type *E. coli* and *Klebsiella pneumoniae*, with a minimum inhibitory concentration (MIC) of 5.6 μg ml^{-1}. As these quinolines represent the first potent and selective inhibitors of MsbA, we sought to understand the molecular basis for their antagonism through structural studies.

Structure of MsbA in complex with G907

Obtaining high-resolution structural information for MsbA has proven challenging[10], and recent cryo-EM structures did not achieve the resolutions necessary to support structure-based inhibitor design[9]. We obtained crystals of *Ec*MsbA in the presence of quinoline inhibitors that diffracted X-rays to 2.9 Å resolution when a facial amphiphile was used[11] (Extended Data Table 1). In complex with G907, MsbA adopts an inward-facing conformation in which a membrane-exposed pocket is formed between transmembrane helices TM4, TM5 and TM6 of each subunit (Fig. 2a–e). G907 is wedged into this crevice in each MsbA

[1]Structural Biology, Genentech Inc., San Francisco, CA, USA. [2]Biochemical and Cellular Pharmacology, Genentech Inc., San Francisco, CA, USA. [3]Infectious Diseases, Genentech Inc., San Francisco, CA, USA. [4]Protein Analytical Chemistry, Genentech Inc., San Francisco, CA, USA. [5]Biomolecular Resources, Genentech Inc., San Francisco, CA, USA. [6]Pathology, Genentech Inc., San Francisco, CA, USA. [7]Discovery Chemistry, Genentech Inc., San Francisco, CA, USA. [8]WuXi Apptec. Co. Ltd., Shanghai, China. [9]These authors contributed equally to this work: Anh Miu, Mary Kate Alexander, Natalie K. Garcia. [10]These authors jointly supervised this work: Jian Payandeh, Christopher M. Koth. *e-mail: payandeh.jian@gene.com; koth.christopher@gene.com

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

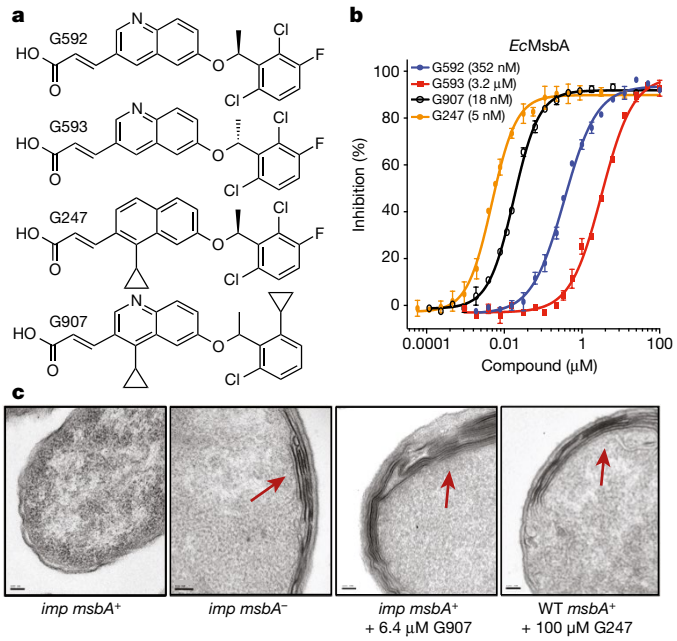


Fig. 1 | Characterization of selective MsbA inhibitors. a, Inhibitors described in this study. **b**, Dose–response curves of purified *E. coli* MsbA in amphipols. Data are mean ± s.e.m. from three independent experiments in all cases except for G593 where data are from two experiments. IC_{50} values (shown in parentheses) were determined by fitting inhibition dose–response curves with a nonlinear four-parameter inhibition model (see Methods section 'MsbA ATPase assay'). **c**, Representative thin section electron micrographs comparing *msbA*$^+$, *msbA*$^-$ and inhibitor-exposed cells, in which *imp* indicates use of the *E. coli* CFT073 UPEC *lptD(imp4213)* strain (Supplementary Table 1). Images are representative of 20 isolated cells from two independent experiments. Scale bars, 0.1 μm.

subunit, halfway across the transmembrane region (Fig. 2a). There is no obvious access pathway to this site from the cytoplasm or inner vestibule, so it appears that G907 gains access to MsbA through the bulk membrane. *Ec*MsbA crystallized in the presence of a second quinoline inhibitor, G092, revealed a similar structure despite different lattice packing (Extended Data Fig. 3a–h and Extended Data Table 1).

The 2-chloro-6-cyclopropylphenyl substituent of G907 (A-ring) exhibits strong electron density (Fig. 2b) and makes van der Waals interactions with side chains from TM4 (L171, A175 and V178), TM5 (A259 and L263) and TM6 (M291 and L294) (Fig. 2d, e and Extended Data Fig. 4). The quinoline core of G907 (B-ring) is orthogonal to the plane of the phenyl substituent (Fig. 2d), where it is partially enclosed by residues from TM4 (V178, S179 and I182), TM5 (A259) and TM6 (M295 and L298) (Fig. 2d, e). A259 and M295 side chains contact the 4-cyclopropyl substitution of the quinoline core and delineate the inhibitor binding-pocket from the inner vestibule of MsbA (Fig. 2d, e). The acrylic acid substitution forms an anchoring salt bridge with R190 (TM5) and points towards the inner vestibule, where the *trans* olefin linker is surrounded by TM4 (I182), TM5 (I255) and TM6 (K299) (Fig. 2d). The tight fit of G907 (and G092) observed within the binding pocket explains enantiomeric selectivity and rationalizes on-target resistance mechanisms for this new class of antibacterial compounds, in which two isolated resistant mutants, A175V and I182M, would sterically interfere with G907 binding (Fig. 2c, Extended Data Fig. 2d, e and Supplementary Information Section 1). As treatment of *E. coli* with quinoline inhibitors causes target-mediated cell killing (Fig. 1c), the G907–LPS–*Ec*MsbA complex probably represents the physiologically relevant inhibitor–MsbA complex in vivo and reveals a previously unknown modulatory site within a B-family ABC transporter.

State-dependent inhibition of MsbA by G907

The outward-facing structure of *Salmonella typhimurium* MsbA (*St*MsbA)[10], and the recently determined inward-facing structure of *Ec*MsbA in complex with LPS[9] (LPS–*Ec*MsbA-cryo), allow direct comparison of the G907-binding site under distinct conformational states (Fig. 3a). In the outward-facing *St*MsbA structure, the G907 pocket is markedly deformed and expected to perturb inhibitor binding (Fig. 3b). This substantial perturbation of the binding site during the transport cycle defines G907 as an inward-facing state-dependent inhibitor. Further comparison with the LPS–*Ec*MsbA-cryo structure suggests an induced fit, whereby G907 binding is accommodated by a shift of TM4 by approximately 4 Å to expose the full binding pocket (Fig. 3c). The presence of G907 causes strain along TM4, as evidenced

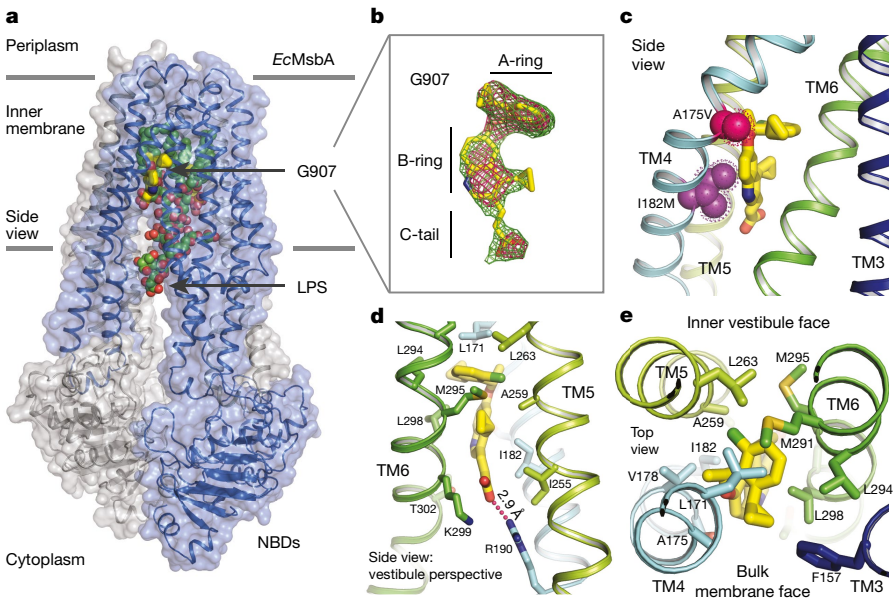


Fig. 2 | The G907–LPS–*Ec*MsbA complex reveals a transmembrane inhibitor binding site. a, Inward-facing conformation of G907–LPS–*Ec*MsbA complex with membrane boundaries, G907 (yellow) and LPS (green) indicated. **b**, $2F_o - F_c$ (1σ, green) and $F_o - F_c$ (2σ, pink) maps calculated before modelling G907, carved at 5 Å. **c**, On-target resistance mutations: A175V (pink) and I182M (purple). **d**, **e**, Alternate inhibitor binding-site views coloured from blue (TM3) to green (TM6).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

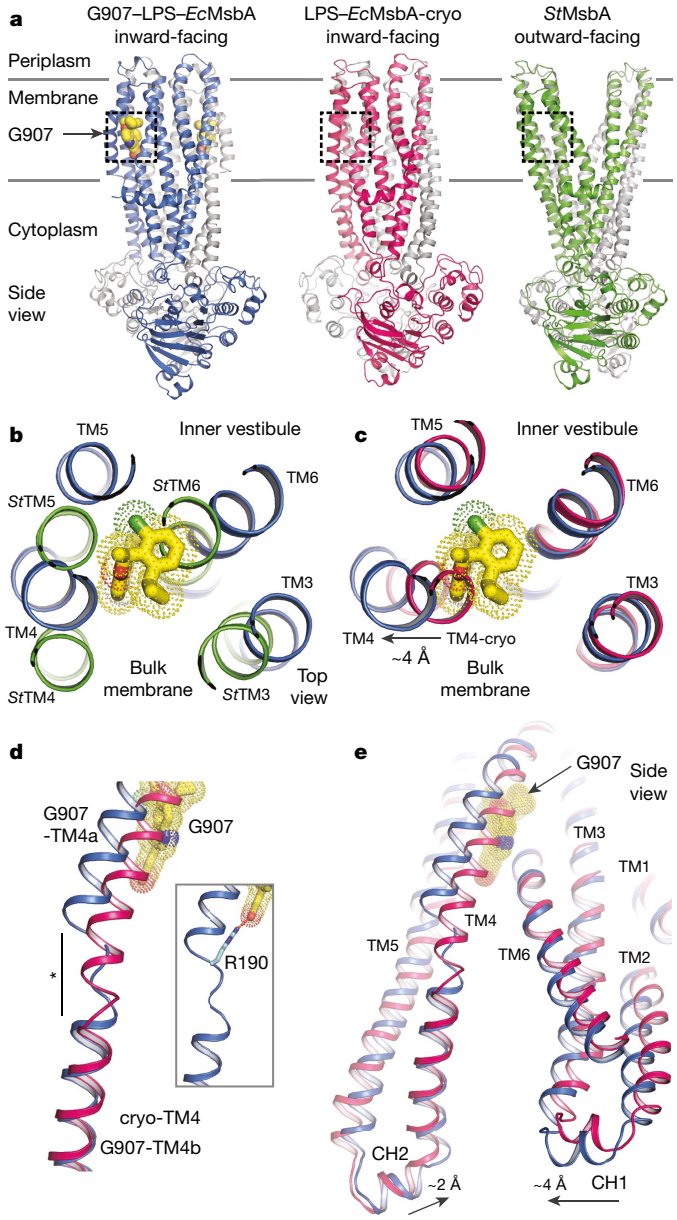


Fig. 3 | Induced fit and state-dependent inhibition of MsbA by G907. **a**, Matched perspectives of G907–LPS–*Ec*MsbA (this study), inhibitor-free LPS–*Ec*MsbA-cryo (Protein Data Bank (PDB) accession 5TV4), and outward-facing *St*MsbA (PDB accession 3B60) complexes; LPS has been omitted for clarity. **b**, Structural superposition reveals deformation of the binding site in the *St*MsbA outward-facing state, where TM5 and TM6 clash with G907. **c**, **d**, Superposition with the inhibitor-free LPS–*Ec*MsbA-cryo structure (pink, cryo-TM4) indicates that part of TM4 translates (blue, G907-TM4a) to form the G907 binding site. TM4 also unwinds (asterisk, G907-TM4b) and R190 forms a salt bridge with the acrylic-acid substitution of G907 (**d**, inset). **e**, Superposition highlights structural differences propagated towards the NBDs, relative to the inhibitor-free LPS–*Ec*MsbA-cryo structure; NBDs and subunit B have been removed for clarity.

by a local unwinding of the TM4 helix below the binding site (Fig. 3d). As a consequence, the R190 side chain becomes available to form an anchoring salt-bridge interaction with the acrylic acid substitution of G907 (Fig. 3d, inset). Overall, these observations are consistent with G907 acting as a selective inhibitor of an inward-facing state of MsbA (Fig. 3b, c). As LPS co-purifies with MsbA (Extended Data Fig. 1i), and G907 is not competitive with LPS (Fig. 2a), G907 appears to act by impeding the structural rearrangements necessary to transition LPS–MsbA to an outward facing conformation (Fig. 3a–c). Notably, G907 targets a site that is critical for mechanical coupling during the transport cycle[12] and architecturally conserved in B-family ABC transporters[13] (Extended Data Fig. 5).

Allosteric inhibition of MsbA by G907

The G907–LPS–*Ec*MsbA structure reveals an unanticipated second mechanism of small-molecule antagonism whereby the two-fold symmetry within the transmembrane region breaks down at the NBDs (Fig. 4a–c). Comparison of pseudosymmetric mock MsbA homodimers generated from the G907–LPS–*Ec*MsbA structure reveals that the NBDs in the A′–A model are separated by approximately 10 Å (Fig. 4b), whereas a severe clash between the NBDs occurs in the B–B′ model (Fig. 4c). This rationalizes why the NBDs in the G907–MsbA crystal structure have adopted an asymmetric conformation: they cannot simultaneously occupy the same physical space. Structurally, coupling helices 1 and 2 (CH1 and CH2) from subunit A have become unlatched from a highly conserved CH1–CH2–NBD interaction network linking the transmembrane domains to the NBDs, as first defined in *Caenorhabditis elegans* P-gp[13] (*Ce*Pgp; Fig. 4d). The pronounced asymmetric arrangement of NBDs in the G907–LPS–*Ec*MsbA complex is expected to be incompetent for ATP hydrolysis (Fig. 4a–d and Extended Data Fig. 6), consistent with the potent inhibition of ATPase activity observed in vitro (Fig. 1b). NBD uncoupling was also observed in solution, as revealed by hydroxyl-radical footprinting–mass spectrometry (HRF–MS)[14]. Relative to the absence of inhibitor, MsbA footprinting analysis in the presence of G907 and G092 revealed a marked increase in the solvent exposure of CH1 and CH2 (Fig. 4e, Extended Data Fig. 7 and Extended Data Table 2).

Although functional asymmetry has been reported for transporters with non-equivalent NBDs[15–18], the structural asymmetry in the G907–LPS–*Ec*MsbA complex is, to our knowledge, unlike that of any other ABC transporter structure yet determined[9,13,15,19] (Fig. 4a–d and Extended Data Fig. 6). The profound uncoupling of the NBD machinery in *Ec*MsbA suggests that G907 stabilizes a catalytically incompetent state of the transporter. Comparison of related B-family ABC transporter structures[9,13,19] reveals that there is a unique displacement of one NBD by 10–15 Å towards the membrane normal in the G907–LPS–*Ec*MsbA complex (Extended Data Fig. 6). The molecular events leading to this NBD displacement originate approximately 50 Å away at the inhibitor-binding site. In comparison to the LPS–*Ec*MsbA-cryo structure[9], inhibitor binding induces movement and strain along TM4 that, through an inward scissoring motion of the TM2–TM3 and TM4–TM5 helices, forces the NBDs together (Fig. 3d, e). Concomitant release of one of the CH1–CH2–NBD interaction networks within the MsbA homodimer is necessary to avoid direct steric clashes between the NBDs (Fig. 4a–d). We conclude that G907 functions as an allosteric dual-mode inhibitor because NBD uncoupling occurs approximately 50 Å away from the primary membrane-embedded binding sites (Fig. 2a).

MsbA uses a hydrocarbon ruler to measure LPS

Until recently, the molecular basis of lipid recognition by MsbA was unknown. Compared to the LPS–*Ec*MsbA-cryo structure[9], a more complete LPS molecule is delineated in the 2.9 Å G907–LPS–*Ec*MsbA and G092–LPS–*Ec*MsbA crystal structures (Fig. 5a, b and Extended Data Fig. 8). Although unambiguous placement of the core sugar moieties is challenging, the characteristic phosphoglucosamine substituents of the lipid A moiety are well-defined and surrounded by a coordination network of arginine side chains arising from the two-fold symmetric transmembrane architecture of the MsbA transporter, approximately halfway across the membrane bilayer (Fig. 5a–c and Extended Data Fig. 8). All six acyl chains of LPS have now been resolved, allowing direct visualization of densely packed hydrocarbon chains that point towards the closed periplasmic gate of MsbA (Fig. 5a–c and Extended Data Fig. 8). The G092–LPS–*Ec*MsbA ternary complex reveals that the inner vestibule of *Ec*MsbA forms a thimble-like structure that is

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

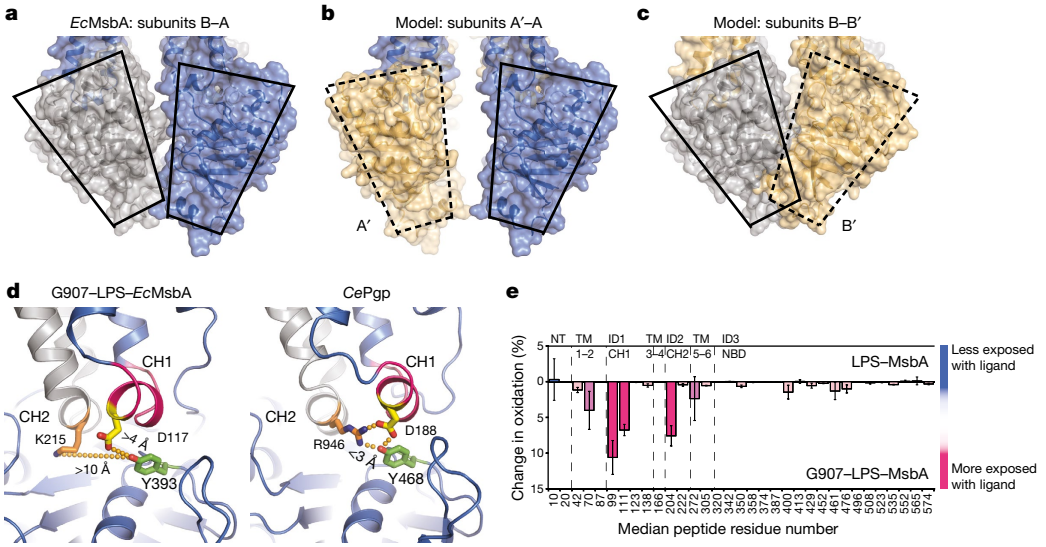


Fig. 4 | Asymmetric uncoupling of the NBDs in MsbA by G907. **a**, NBDs of G907–LPS–*Ec*MsbA (subunit A, blue; subunit B, grey). **b**, **c**, Pseudosymmetric A′–A (**b**) and B–B′ (**c**) models built by superimposing the transmembrane domains. **d**, NBD–CH1–CH2 coupling interface. *Ec*MsbA residues D117 (yellow), K215 (orange) and Y393 (green) are highly conserved amongst B-family ABC exporters, as in *Ce*Pgp (PDB accession 4F4C), in which these residues typically form coupling interactions (<3.5 Å). **e**, Solution-based HRF–MS performed on LPS–*Ec*MsbA with or without G907. Tryptic peptide oxidation of inhibitor-bound sample relative to inhibitor-free sample is shown. Data are mean ± s.e.m. from three replicate experiments. NT, N terminus; ID1, ID2 and ID3 refer to the intracellular (cytoplasmic) CH1, CH2 and NBD regions of MsbA, respectively.

expected to function as a hydrocarbon ruler to select for the 12-carbon and 14-carbon chains characteristic of LPS (Fig. 5a and Extended Data Fig. 8). As it would be less favourable to constrain the longer 18-carbon chains of a typical glycerophospholipid between the charged arginine coordination ring and the closed periplasmic gate (Fig. 5a), *Ec*MsbA targets this second distinguishing feature of lipid A to achieve substrate selectivity over competing bulk membrane phospholipids, in addition to securing the bivalent phosphoglucosamine headgroup.

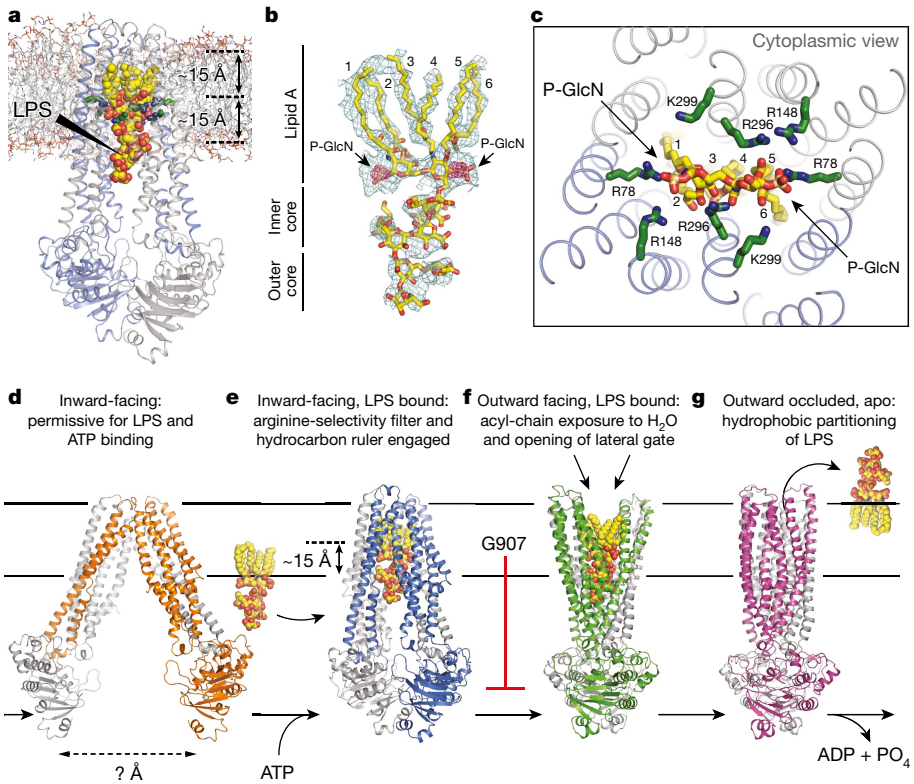


Fig. 5 | LPS selectivity and transport by MsbA. **a**, G092–LPS–*Ec*MsbA complex highlighting LPS (yellow) and arginine and lysine residues (green) involved in lipid A coordination; G092 omitted. Bulk membrane is depicted on the basis of a molecular dynamics simulation. **b**, $F_o - F_c$ maps calculated with LPS omitted (1.0σ, cyan; 5.0σ, pink), carved at 5 Å. **c**, Basic residues (green) form a selectivity filter around the phospho-*N*-acetylglucosamine (P-GlcN) moieties of lipid A, halfway across the membrane (see **a** for perspective). **d**–**g**, LPS transport model; see text for details. **d**, *E. coli* MsbA (PDB accession 3B5W); **e**, G907–LPS–*Ec*MsbA as a proxy for the LPS-bound inward-facing state; **f**, *St*MsbA (PDB accession 3B60); **g**, ADP–vanadate–*Ec*MsbA-cryo (PDB accession 5TTP).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Implications for MsbA and ABC transporters

The dual-mode inhibitory mechanism of G907 uncovered here may have broad but unrecognized relevance across the ABC transporter superfamily. A pocket analogous to the G907-binding site has been observed in related ABC transporters[13] (Extended Data Fig. 5), so it is tempting to speculate that this site may be an allosteric modulatory hotspot for natural ligands, lipids, protein partners or designed small molecules. Furthermore, although the site and mechanisms by which G907 antagonizes MsbA are clearly distinct, it is noteworthy that the antagonist glibenclamide also wedges between transmembrane helices in the ABC transporter-like protein SUR1[20], and that a nanobody inhibitor of PglK sterically impairs NBD closure, ATP hydrolysis and transition to an outward-facing conformation[21]. The SUR1 and PglK structures therefore independently highlight two emerging modulatory concepts discovered for the dual-mode quinoline series inhibitors targeting MsbA: transmembrane-helical interdigitation and functional NBD uncoupling.

The structure of LPS in complex with *Ec*MsbA presented herein helps to reconcile evolutionarily conserved features of the LPS biosynthesis pathway[4]. Overexpression of *msbA* suppresses mutations in the *htrB* lipid A acyltransferase, as MsbA can transport the tetra-acylated LPS produced in *htrB* mutants, albeit inefficiently[6]. The tight packing of hexa-acylated LPS observed in our crystal structures (Fig. 5a, b and Extended Data Fig. 8) suggests that the binding of acyl chains into the inner vestibule of MsbA provides an element of substrate selectivity, rationalizing why tetra-acylated substrates are less preferred. MsbA cannot, however, efficiently transport a substrate lacking phosphorylation at the 4′-position of lipid A. This last essential step of lipid A biosynthesis is catalysed by LpxK[22,23]. Notably, the *lpxK* gene is found immediately downstream of *msbA* in *E. coli*, and the two proteins are concatenated in some bacteria, suggesting tight coupling of their cellular activities. MsbA has evolved to directly target this unique feature of LPS as a primary basis for lipid selectivity (Fig. 5c). Although the overall features revealed by our structures do not rule out the possibility that other lipids are transported by MsbA, the multi-point coordination of the bivalent lipid A phosphoglucosamine moieties by an arginine selectivity filter and the snug fit of the short acyl chains within the inner vestibule are features that are consistent with a preference for selective LPS transport.

The G907–LPS–*Ec*MsbA and G092–LPS–*Ec*MsbA complexes, in combination with structural information from other ABC transporters[9,10], define an alternating access mechanism of LPS transport that is summarized in Fig. 5d–g. An open inward-facing conformation of MsbA must exist that promotes lateral ferrying of the massive LPS substrate from the membrane inner leaflet into the vestibule of MsbA (Fig. 5d, e). Upon binding, MsbA completely engulfs its LPS substrate (Fig. 5a, e), in contrast to other lipid transporters that leave the acyl chains of their substrates within the bulk bilayer[12,24,25]. Preferential binding to the arginine coordination ring shuttles LPS halfway across the bilayer and shields its polar headgroup (Fig. 5e), where an intrinsic hydrocarbon ruler measures acyl chain length against a closed periplasmic gate to further ensure selective substrate engagement. In the absence of a G907-like inhibitor, ATP binding promotes NBD closure and an outward-facing conformation (Fig. 5f). Upon opening of the periplasmic gate (Fig. 5f, g), acyl-chain exposure to the aqueous environment and concomitant opening of a lateral gate provide a feasible pathway for LPS to exit MsbA and become flipped into the outer leaflet of the bilayer (Fig. 5f, g). LPS discharge is probably also aided by ATP hydrolysis[9], and after the release of inorganic phosphate and ADP from the NBDs, MsbA recycles back to an inward-facing conformation (Fig. 5d).

The identification of novel dual-mode MsbA antagonists has allowed us to visualize unprecedented structural features of ABC transporter inhibition and lipid substrate selection. G907 and related quinoline antagonists do not inhibit MsbA by competing directly with the transported substrate, as is typical for other ABC transporter inhibitors. Unexpectedly, binding of inhibitors such as G907 into a transmembrane-exposed pocket promotes a cascade of molecular events that are propagated along TM4 and ultimately result in the unhinging of a conserved interaction network and subsequent functional and structural uncoupling of the NBDs. Considering the architectural conservation of the G907-binding pocket in other ABC transporters, this work should enable the discovery of selective antagonists across the broader ABC transporter superfamily.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0083-5.

Received: 25 July 2017; Accepted: 20 March 2018;
Published online 2 May 2018.

1. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336–343 (2016).
2. Laxminarayan, R. et al. Antibiotic resistance—the need for global solutions. *Lancet Infect. Dis.* **13**, 1057–1098 (2013).
3. Nikaido, H. Molecular basis of bacterial outer membrane permeability revisited. *Microbiol. Mol. Biol. Rev.* **67**, 593–656 (2003).
4. Raetz, C. R. H. & Whitfield, C. Lipopolysaccharide endotoxins. *Annu. Rev. Biochem.* **71**, 635–700 (2002).
5. Sperandeo, P. et al. Characterization of *lptA* and *lptB*, two essential genes implicated in lipopolysaccharide transport to the outer membrane of *Escherichia coli*. *J. Bacteriol.* **189**, 244–253 (2007).
6. Zhou, Z., White, K. A., Polissi, A., Georgopoulos, C. & Raetz, C. R. Function of *Escherichia coli* MsbA, an essential ABC family transporter, in lipid A and phospholipid biosynthesis. *J. Biol. Chem.* **273**, 12466–12475 (1998).
7. Doerrler, W. T., Reedy, M. C. & Raetz, C. R. An *Escherichia coli* mutant defective in lipid export. *J. Biol. Chem.* **276**, 11461–11464 (2001).
8. Davidson, A. L., Dassa, E., Orelle, C. & Chen, J. Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol. Mol. Biol. Rev.* **72**, 317–364 (2008).
9. Mi, W. et al. Structural basis of MsbA-mediated lipopolysaccharide transport. *Nature* **549**, 233–237 (2017).
10. Ward, A., Reyes, C. L., Yu, J., Roth, C. B. & Chang, G. Flexibility in the ABC transporter MsbA: alternating access with a twist. *Proc. Natl Acad. Sci. USA* **104**, 19005–19010 (2007).
11. Lee, S. C. et al. Steroid-based facial amphiphiles for stabilization and crystallization of membrane proteins. *Proc. Natl Acad. Sci. USA* **110**, E1203–E1211 (2013).
12. Locher, K. P. Mechanistic diversity in ATP-binding cassette (ABC) transporters. *Nat. Struct. Mol. Biol.* **23**, 487–493 (2016).
13. Jin, M. S., Oldham, M. L., Zhang, Q. & Chen, J. Crystal structure of the multidrug transporter P-glycoprotein from *Caenorhabditis elegans*. *Nature* **490**, 566–569 (2012).
14. Zhang, Y., Rempel, D. L., Zhang, H. & Gross, M. L. An improved fast photochemical oxidation of proteins (FPOP) platform for protein therapeutics. *J. Am. Soc. Mass Spectrom.* **26**, 526–529 (2015).
15. Esser, L. et al. Structures of the multidrug transporter P-glycoprotein reveal asymmetric ATP binding and the mechanism of polyspecificity. *J. Biol. Chem.* **292**, 446–461 (2017).
16. Gao, M. et al. Comparison of the functional characteristics of the nucleotide binding domains of multidrug resistance protein 1. *J. Biol. Chem.* **275**, 13098–13108 (2000).
17. Procko, E., Ferrin-O'Connell, I., Ng, S. L. & Gaudet, R. Distinct structural and functional properties of the ATPase sites in an asymmetric ABC transporter. *Mol. Cell* **24**, 51–62 (2006).
18. Sorum, B., Töröcsik, B. & Csanády, L. Asymmetry of movements in CFTR's two ATP sites during pore opening serves their distinct functions. *eLife* **6**, e29013 (2017).
19. Nöll, A. et al. Crystal structure and mechanistic basis of a functional homolog of the antigen transporter TAP. *Proc. Natl Acad. Sci. USA* **114**, E438–E447 (2017).
20. Martin, G. M., Kandasamy, B., DiMaio, F., Yoshioka, C. & Shyng, S. L. Anti-diabetic drug binding site in a mammalian K_{ATP} channel revealed by cryo-EM. *eLife* **6**, e31054 (2017).
21. Perez, C. et al. Structural basis of inhibition of lipid-linked oligosaccharide flippase PglK by a conformational nanobody. *Sci. Rep.* **7**, 46641 (2017).
22. Ray, B. L. & Raetz, C. R. The biosynthesis of gram-negative endotoxin. A novel kinase in *Escherichia coli* membranes that incorporates the 4′-phosphate of lipid A. *J. Biol. Chem.* **262**, 1122–1128 (1987).
23. Garrett, T. A., Kadrmas, J. L. & Raetz, C. R. Identification of the gene encoding the *Escherichia coli* lipid A 4′-kinase. Facile phosphorylation of endotoxin analogs with recombinant LpxK. *J. Biol. Chem.* **272**, 21855–21864 (1997).
24. Baldridge, R. D. & Graham, T. R. Identification of residues defining phospholipid flippase substrate specificity of type IV P-type ATPases. *Proc. Natl Acad. Sci. USA* **109**, E290–E298 (2012).
25. Brunner, J. D., Lim, N. K., Schenck, S., Duerst, A. & Dutzler, R. X-ray structure of a calcium-activated TMEM16 lipid scramblase. *Nature* **516**, 207–212 (2014).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Acknowledgements We thank our colleagues for support, particularly E. Brown, S. Hymowitz, B. Roth, M. Starovasnik and W. Young. We also thank V. Janakiraman, M. McCreary, E. Skippington, and K. Toy for whole-genome sequencing and analysis. Studies were carried out at the Stanford Synchrotron Radiation Lightsource 12-12 at Stanford Linear Accelerator Center National Accelerator Laboratory; use of this facility is supported by the US Department of Energy (DOE), DOE Office of Biological and Environmental Research, National Institutes of Health (NIH), and National Institute of General Medical Sciences (NIGMS). Studies were also carried out at the Berkeley Center for Structural Biology, supported by the NIH, NIGMS, and the Howard Hughes Medical Institute. The Advanced Light Source is a DOE Office of Science User Facility, contract DE-AC02-05CH11231. Reagents are available under a materials transfer agreement with Genentech.

Reviewer information *Nature* thanks E. Breukink, P. Nissen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions H.Ho. and A.O. purified and reconstituted MsbA. H.Ho. crystallized MsbA complexes. H.Ho., A.M., A.O., J.Q., K.C., M.H.B., T.M. and Y.X. performed high-throughput screening and biochemical studies. M.K.A. and M.N. performed phenotypic studies. N.K.G. and A.T.W. performed HRF–MS experiments. H.Ho., I.Z., C.T., S.S. and Y.F. performed molecular biology and protein expression. M.R. and C.D.A. performed electron micrograph studies. H.Hu., S.S.L., J.L., L.W., J.W., Y.L., H.E.P., B.D.S., M.F.T.K., and V.V. performed chemical designs and syntheses. A.M., M.K.A., N.K.G., A.O., M.R., H.E.P., M.H.B., M.-W.T., B.D.S., J.R.K., A.T.W., V.V., Y.X., M.N., J.P. and C.M.K. analysed the data; B.D.S. performed MD simulations, J.P. determined X-ray structures with input from J.R.K.; J.P. and C.M.K. wrote the manuscript with input from all authors; J.P. and C.M.K. supervised the project and are co-senior authors.

Competing interests All authors except J.W. and Y.L. are employees of Genentech; J.W. and Y.L. are employees of WuXi.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0083-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0083-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.M.K. and J.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Protein expression and purification. The wild-type MsbA transporter from *E. coli* was cloned into a modified pET-52b expression vector and overexpressed using *E. coli* host Rosetta 2 (DE3) grown by fermentation at 17 °C for 64 h. Cells expressing *E. coli* MsbA were collected and resuspended in 50 mM Tris, pH 8.0, 500 mM NaCl (Buffer A) supplemented with cOmplete Protease Inhibitors (Roche), 1 mM PMSF and 2 U ml^{-1} of Benzonase nuclease (Sigma Aldrich). After cell lysis by microfluidization, FA-3 detergent (for structural studies and some biochemical assays, as indicated; Avanti Polar Lipids) or lauryl maltose neopentyl glycol (LMNG, for biochemical assays; Anatrace) was added to 1% (w/v) and protein solubilization was carried out with gentle agitation at 4 °C overnight. After centrifugation at 185,000*g* for 1 h, clarified supernatant was mixed gently with anti-Flag M2-agarose resin (Sigma Aldrich) pre-equilibrated with buffer B (buffer A supplemented with 0.06% (w/v) FA-3 or 0.005% LMNG (w/v)) for 1–2 h at 4 °C. Flag resin was collected by gravity flow, washed with ten column volumes of buffer B, and eluted with four column volumes of buffer B supplemented with 0.15 mg ml^{-1} Flag peptide. The Flag-tag was removed using TEV protease and the cleaved *E. coli* MsbA was passed over a Superdex 200 column (GE Healthcare) in 20 mM Tris pH 8, 100 mM NaCl and 0.06% (w/v) FA-3 or 0.005% LMNG (w/v). The peak fractions containing *E. coli* MsbA were pooled and concentrated to 10 mg ml^{-1} using Vivaspin centrifugal devices (50 K MWKO). Protein aliquots were flash frozen and stored at −80 °C.

MsbA reconstitution into amphipols. Where the use of amphipol-incorporated MsbA is indicated, the purified MsbA in LMNG was adjusted to 1 mg ml^{-1} and reconstituted into A8-35 amphipols[26](Anatrace) by addition of an equal volume amphipol stock (20 mg ml^{-1}). The sample was incubated overnight at 4 °C and resolved again over a Superdex 200 column (GE Healthcare) in 100 mM NaCl, 20 mM Tris pH 8.0. The peak fractions containing *E. coli* MsbA were pooled and concentrated to 1 mg ml^{-1} using Vivaspin centrifugal devices (50 K MWKO). Protein aliquots were flash frozen and stored at −80 °C.

Quantitation of co-purifying LPS. The LPS content of purified *E. coli* MsbA in amphipol and FA-3 incorporated MsbA was measured using a *Limulus* amoebocyte lysate chromogenic endotoxin quantitation assay, according to the manufacturer's instructions (Pierce). A standard curve was generated using LPS from *E. coli* strain O111:B4, as directed by the manufacturer. One endotoxin unit (EU) was assumed to be equal 0.1 ng of LPS.

Reconstitution into nanodiscs. The process of *E. coli* MsbA reconstitution into nanodiscs[27] was initiated by mixing LMNG-solubilized MsbA protein with purified recombinant membrane scaffold protein (MSP) and sodium cholate-solubilized phospholipids 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoethanolamine (POPE):phosphatidylglycerol (POPG) (9:1 w/w ratio). The sample mixture was dialysed into 100 mM NaCl, 20 mM Tris, pH 8.0 overnight at room temperature. *E. coli* MsbA nanodiscs were separated from protein aggregates and free MSP by resolving over a Superdex 200 column (GE Healthcare). Fractions containing *E. coli* MsbA nanodiscs were pooled, concentrated to 1 mg ml^{-1} and stored at −80 °C.

MsbA ATPase assay. The ATPase activity of MsbA was measured using a Tanscreener ADP2 FP Assay (BellBrook Labs). To determine the IC_{50} of MsbA inhibitors, compounds were incubated with 2× MsbA enzyme solution for 10 min and then 2× ATP solution was added to initiate the ATPase reaction. The final condition of the ATPase reaction was 3 nM MsbA (*E. coli* or *Klebsiella*, as indicated) in amphipol, 50 μM ATP in 50 mM Tris pH 7.5, 10 mM $MgCl_2$, 1% glycerol, 0.1% bovine gamma globulin, 1 mM dithiothreitol (DTT), 0.007% Brij-35 and 0.5% dimethylsulfoxide (DMSO). The A175V MsbA mutant was screened at a final concentration of 9 nM. The reaction was quenched by the addition of Transcreener Detection Buffer when about 5% of the ATP was hydrolysed to ADP. The IC_{50} was determined by fitting the inhibition dose–response curve with a nonlinear four-parameter inhibition model (GraphPad Prism). The same assay conditions were used in the uHTS campaign (using *E. coli* MsbA in nanodiscs).

Bacterial strains. Most *E. coli* strains used in this study were constructed in CFT073, an uropathogenic *E. coli* (UPEC) strain (Supplementary Table 1). Strain identity was confirmed by whole genome sequencing. For MsbA overexpression, the *msbA* open reading frame (ORF) was cloned into the EcoRI/XbaI sites of the arabinose-inducible, high-copy expression vector pBAD24 using PCR primers *msbA*-f2 and *msbA*-r2 (Supplementary Table 2). Overexpression of the MsbA protein was confirmed by western blotting using polyclonal antibodies (1:2,000 sera dilution) raised against purified MsbA protein (Extended Data Fig. 2a).

A conditional MsbA knockout (*msbA*-cond-ko) strain was constructed by subcloning the arabinose-inducible *msbA* from pBAD24 into the SacI site of the integration vector pLDR9[28] using PCR primers PBAD-*msbA*-f and PBAD-*msbA*-r (Supplementary Table 2). The NotI-digested and religated construct was integrated into the lambda *att* site using lambda integrase[28]. The endogenous *msbA* gene was then replaced with a promoterless kanamycin marker flanked by FRT sites, amplified from pKD4 using primers *msbA*-KO-noPkan-f and *msbA*-KO-kan-r (Supplementary Table 2) and integrated into CFT073 by Red recombinase-mediated homologous recombination[29].

To make strains with a permeablized outer membrane, the endogenous *lptD* gene of the wild-type CFT073 or CFT073 *msbA*-cond-ko strain was replaced with an *lptD*(*imp4213*) mutant allele[30]. A construct was synthesized (GenScript) with 200 bp of the *lptD* ORF just upstream of the *lptD*(*imp4213*) deletion followed by the remaining 1,299 bp of the *lptD* ORF codon changed to keep the same amino acid sequence but prevent recombination during integration, a promoterless kanamycin resistance marker flanked by FRT sites, and 200 bp homologous to the chromosomal sequence just downstream of the *lptD* ORF. The construct was integrated onto the chromosome by Red recombinase-mediated homologous recombination[28].

An efflux-deficient mutant of the *E. coli* K12 strain MG1655 (ATCC 700926), MG1655 Δ*tolC*, was made by transducing the Δ*tolC*::*kanR* allele from the Keio collection strain JW5503[31] using P1*vir* phage (*E. coli* Genetic Stock Center 12133).

Cellular IC_{50} and MIC determinations. Two-hundred-nanolitre aliquots of eight twofold dilutions of each compound were dispensed in duplicate into 384-well plates using an Echo Liquid Handler (Labcyte). Strains to be tested were grown in M9 medium supplemented with 1% maltose and, for the overexpression strains, 0.2% arabinose overnight for 16 h, washed with PBS, resuspended at $OD_{600} = 1$, and frozen in 100-μl aliquots for use in screening. For each experiment, aliquots were thawed on ice and diluted 1:1,000 into M9 1% maltose, plus 0.2% arabinose for overexpression strains. Ten microlitres of cells per well was dispensed into 384-well plates containing compound using a Multidrop Combi Reagent Dispenser (Thermo). Plates were incubated for 6 h at 37 °C, then 10 μl BacTiter-Glo reagent (Promega) was added to each well and luminescence was measured. IC_{50} values were determined by fitting the inhibition dose–response curve with a nonlinear four-parameter inhibition model (GraphPad Prism). Reported values are the standard geometric means of ≥2 independent values. MICs were determined in cation-adjusted Mueller–Hinton broth with 0.002% Tween-80 following the standard Clinical and Laboratory Standards Institute protocol[32].

Electron microscopy. To prepare cells for electron microscopy, an overnight culture of CFT073 *lptD(imp4213)* grown in LB was diluted 1:100 in LB; in addition an overnight culture of CFT073 *lptD(imp4213) msbA*-cond-ko, which was grown in LB 2% arabinose to induce MsbA expression to wild-type levels, was diluted 1:100 in LB 2% arabinose. Both cultures were grown to log phase for 2.5 h at 37 °C, then pelleted, washed twice with PBS and resuspended at $OD_{600} = 0.1$ in LB. Fifty microlitres of 10 mM G03047592 or 50 μl DMSO (no drug control) was added to 5 ml CFT073 *lptD(imp4213)*; 1% glucose was added to a 5-ml culture of CFT073 *lptD(imp4213) msbA*-cond-ko to fully repress MsbA expression. All cultures were incubated for 3 h at 37 °C. Cells were pelleted, washed with PBS, fixed with 100 μl Karnovsky's fixative (2% paraformaldehyde and 2.5% glutaraldehyde in 0.1 M sodium cacodylate buffer, pH 7.2), post-fixed in 1% reduced osmium tetroxide (EM Sciences) for 1 h and then incubated in 0.5% uranyl acetate for 1 h. The samples were dehydrated through an ascending series of ethanol (50%, 70%, 90%, 100%), then washed with propylene oxide and finally embedded in Eponate 12 (Ted Pella). Ultrathin sections (80 nm) were cut with an Ultracut microtome (Leica), stained with 0.2% lead citrate and examined in a JEOL JEM-1400 transmission electron microscope at 80 kV. Digital images were captured with a GATAN Ultrascan 1000 CCD camera.

Isolation of resistant mutants and whole genome sequencing. Resistant isolates to G592 were obtained by plating MG1655 Δ*tolC* on LB agar with 1.8 μg ml^{-1} G592 (2 × MIC). Growth of MG1655 Δ*tolC* at 4 × MIC gave rise to no viable colonies. Genomic DNA was isolated from 2 ml of overnight culture of the resistant mutants using a DNEasy Blood and Tissue Kit (Qiagen). Genomic DNA was sheared between 350 and 500 bp using a Covaris LE220. The libraries were prepared using a TruSeq Nano library prep kit from Illumina (FC-121-4001). Sequencing was carried out in the Illumina Miseq (1× 150 cycles). An average of ~21 million 2 × 75-bp paired reads were obtained per sample. All fastq files were then filtered for read quality, keeping reads in which at least 70% of the cycles had Phred scores ≥ 23. Filtered reads were aligned to the *E. coli* K-12 MG1655 reference genome (NCBI Reference Sequence accession number NC_000913) using GSNAP version 2013-10-10[33] and variant analysis was performed using the Bioconductor package VariantTools[34]. To increase confidence in variant calls, base quality scores of 30 or higher were required.

Protein crystallization. Inhibitor G907 (or G092) was added to *E. coli* MsbA at 10 mg ml^{-1} in FA-3 to a final concentration of 1 mM and incubated for 1 h on ice. The G907–LPS–*Ec*MsbA (or G092–LPS–*Ec*MsbA) complex was crystallized at 19 °C using sitting-drop vapour diffusion by mixing the complex with mother liquor (200 mM NaCl, 19% PEG 550 MME, 4% PEG 400, 100 mM HEPES, pH 7.0) at a 1:1 (v/v) ratio. Crystals reached full size after 1 week and were then flash frozen in cryoprotectant containing crystallization buffer plus additional 20% PEG 400.

Heavy atom labelling for crystallographic phasing. *E. coli* MsbA purified in FA-3 detergent was adjusted to 1 mg ml^{-1}. Ethyl mercurithiosalicyclic sodium salt (Hg)

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

was added to the final concentration of 1 mM and the mixture was incubated overnight at 4 °C. Free mercury was removed from Hg-labelled MsbA protein by passing the sample onto a Superdex 200 column in 100 mM NaCl, 0.06% FA-3, 20 mM Tris, pH 8.0. The fractions containing Hg-labelled MsbA were pooled and concentrated to 10 mg ml^{-1} using Vivaspin (50 K MWCO) centrifugal devices. MsbA inhibitor G907 was added to the final concentration of 1 mM and the mixture was incubated on ice for 1 h. Crystal trays were set up by mixing MsbA protein sample at 1:1 (v/v) ratio with mother liquor containing 200 mM NaCl, 20% PEG 550 MME, 100 mM HEPES, pH 7.0. Crystals were grown at 19 °C in sitting drops for one week then collected and flash frozen in cryoprotectant containing crystallization buffer with increased PEG 500 MME concentration at 30%.

Structure determination, refinement and analysis. Diffraction data were collected at 100 K using beamline 12-2 of the Stanford Synchrotron Radiation Lightsource and beamline 5.0.2 of the Advanced Light Source. All X-ray diffraction data were integrated and scaled with the HKL2000 suite[35] and further processed with the CCP4 package[36]. All attempts to determine the G907–LPS–*Ec*MsbA structure by molecular replacement using a wide range of available ABC transporter models (full and partial models) failed. Because wild-type *Ec*MsbA contains two native cysteine residues (Cys88 and Cys315) per subunit, we screened and produced a covalent mercury complex of *Ec*MsbA using the fluorescence detection of heavy-atom labelling (FD–HAL) method[37]. This Hg–*Ec*MsbA complex was subsequently co-crystallized with G907 as described above. To obtain experimental phases, a SAD dataset was collected at ALS 5.0.2 near the mercury absorption edge ($\lambda = 1.005$ Å) from a mercury-containing complex of G907–LPS–*Ec*MsbA that was crystallized in the *P*1 space group (Extended Data Table 1). This *P*1 space group was ultimately found to contain two full MsbA transporters within the asymmetric unit. Phenix AutoSol[38] was initially able to identify five of a possible eight Hg sites within the heavy atom substructure at ~8.3 Å (anomalous signal 0.1158; SKEW = 0.023; CORR_RMS = 0.738; FOM = 0.285), and according to standard phasing and density modification protocols (FOM = 0.73), all 12 transmembrane helices of the transporter could be clearly resolved using data extending to ~3.7 Å resolution. The transmembrane skeleton of MsbA was manually built initially as a poly-Ala structure. For the NBDs, an available high-resolution model of the *Haemophilus influenzae* MsbA NBD (PDB accession 5EUM) was truncated to poly-Ala and placed manually into the observable NBD density. Reiterative rounds of molecular replacement with SAD phasing (MR–SAD) using PHENIX[38] were employed to extend the backbone structure (for example, periplasmic loops) and make observable rigid body adjustments within the NBDs. Relative to the transmembrane domains of *Ec*MsbA, the structural asymmetry of the NBDs was apparent at very early stages of mercury-phased model building. A 2.9-Å resolution native dataset from a G907–LPS–*Ec*MsbA crystal in space group $P2_1$ was obtained, collected ($\lambda = 1.000$ Å) from SSRL 12-2, and found to contain one MsbA transporter (homodimer) within the asymmetric unit. Molecular replacement using an early model of G907–LPS–*Ec*MsbA was performed using PHENIX software. Reiterative rounds of refinement and model building in Coot[39] were guided by inspection of omit maps. Strict geometry and secondary structure restraints were applied throughout refinement to maintain stringent stereochemistry; however, non-crystallographic symmetry restraints were ultimately not applied because of the asymmetry within the NBDs. Although strong omit density was apparent at very early stages, coordinates for G907, LPS and phospholipids (assigned as phosphatidylethanolamine bound along the bulk membrane face of *Ec*MsbA) were added only at very late stages of refinement. A 2.92 Å resolution native dataset from a G092–LPS–*Ec*MsbA crystal in space group $C2_1$ was also obtained, collected ($\lambda = 1.000$ Å) from SSRL 12-2, and also found to contain one MsbA transporter (homodimer) within the asymmetric unit. As a starting point, all ligands from the refined G907–LPS–*Ec*MsbA structure were removed and the B-factors of the remaining model were reset to 50 Å^2 and then molecular replacement was performed using PHENIX software. Initially, real-space rigid-body adjustments of the NBDs and transmembrane helices were performed in Coot; then reiterative rounds of refinement with strict geometry and secondary structure restraints applied throughout were guided by inspection of omit maps in Coot[39]. Coordinates for G092, LPS and phospholipids (assigned as phosphatidylethanolamine) were added only at very late stages of refinement. The geometry of the final G907–LPS–*Ec*MsbA and G092–LPS–*Ec*MsbA structural models was assessed using MolProbity[40] and structural alignments were performed using LSQMAN[41]. All structural figures were prepared with the PyMol software[42].

HRF–MS experiments. *E. coli* MsbA at 10 mg ml^{-1} in FA-3 detergent was diluted to 0.9 mg ml^{-1} in 100 mM NaCl, 0.06% FA-3, 20 mM Tris pH 8.0. Three samples were prepared for FPOP analysis: inhibitor-free MsbA and MsbA incubated with 1 mM of either G907 or G092, after lyophilization of the appropriate amount of each inhibitor in the sample tube. HRF–MS analysis was performed using fast photochemical oxidation of proteins (FPOP) as previously described[43]. To avoid light exposure, MsbA solution and 30 mM H_2O_2 were rapidly mixed through a micro-tee mixer (Cobert Associates Laboratory) at flow rates of 15 μl min^{-1} and 7.5μl min^{-1}, respectively. The oxidation reaction was initiated by pulsing a focused 248 nm KrF excimer laser (GAM Laser) at 75 mJ per pulse at a rate of 7 Hz through a 3.0 mm exposure window of 150 μm internal diameter fused-silica tubing (Polymicro Technologies). Oxidized samples were collected in a tube containing a quench solution with methionine (Sigma Aldrich) and catalase (Sigma Aldrich) in a final concentration of 40 mM and 1 μM respectively. Each quenched sample was frozen before proteolysis and mass spectrometry. To account for background oxidation of MsbA, an unexposed sample was obtained and compared to three oxidized replicates for each protein state.

After FPOP, each protein sample was reduced and denatured at 45 °C for 10 min in reducing (10 mM DTT, 25 mM Tris, pH 8.0) and denaturation buffer (6 M guanidinium hydrochloride, 1.0 mM EDTA, 250 mM Tris, pH 8.0). Reduced cysteines were alkylated with 25 mM sodium iodoacetate in denaturation buffer for 5 min at 45 °C and quenched with 50 mM DTT at room temperature for 5 min. The samples were desalted before proteolysis over an Illustra NAP-5 column (GE Healthcare Bio-Sciences) equilibrated in 25 mM Tris, pH 8.0, and the eluate was digested with 5 μg of sequencing grade recombinant Trypsin (Roche) at 37 °C for 1 h. The digest was quenched with formic acid to a final percentage of 3.5 for liquid chromatography.

Ten micrograms of tryptic peptides of MsbA was resolved on a Waters ACQUITY UPLC using a CSH C18 (130 Å,1.7 μm, 2.1 × 150 mm) column heated to 77 °C. Peptides were eluted with a gradient of 99% buffer A (water, 0.1% formic acid) to 50% buffer B (acetonitrile, 0.1% formic acid) over 40 min at 0.3 ml min^{-1}. To prevent carry-over between injections, the column was washed with 95% buffer B for 4 min and re-equilibrated to 99% buffer A for 20 min before the next injection. Peptide spectra were collected on a Thermo Fisher Q-Exactive hybrid quadrupole-orbitrap mass spectrometer. Data-dependent acquisition was used to collect tandem mass spectrometry (MS/MS) spectra of the most abundant peptides for the oxidized and unoxidized samples. Unmodified and oxidized species for each tryptic peptide were identified from the fragmentation spectra using Byonic (Protein Metrics). Percentage oxidation was calculated as previously described using the extracted mass of each identified species from the precursor scans (350–2000 m/z) with Byologic (Protein Metrics)[44,45]. The tryptic peptides measured are summarized in Extended Data Table 2.

Molecular dynamics. To estimate how the MsbA dimer sits relative to the bacterial inner membrane, a complete model of MsbA derived from the refined G907–LPS–*Ec*MsbA 2.9-Å X-ray structure was placed in a simulated lipid and water environment. An all-atom model was generated from the X-ray crystal structure using the Protein Preparation function in Maestro[46,47] which adds missing residues, side chains and hydrogens, predicts residue protonation states, and optimizes side-chain conformations. The G907 inhibitors, phospholipid, and LPS molecules were then removed. A simulation box of protein, lipids, water and ions was generated using the System Builder in Maestro[47] using the following steps. The protein was first placed in a POPE lipid bilayer. The bilayer was initially aligned manually to the region where the protein surface is most hydrophobic. The system was neutralized with the addition of seven chlorine ions. An orthorhombic box was constructed with a 10-Å buffer around the protein in all dimensions and the regions of the box not occupied by protein or lipid were filled with transferable intermolecular potential with three points (TIP3P) waters. The resulting system was then equilibrated using the relax_membrane.py[47] and Desmond multisim[48–50] programs. A final 10-ns NVT simulation was run using Desmond, with a temperature of 300 K, a 2-ps time step, and with the protein backbone constrained (force constant of 5.0 kcal mol^{-1} Å). The lipids from the final frame of the simulation were used and presented in Fig. 5a and Extended Data Fig. 8d–f.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

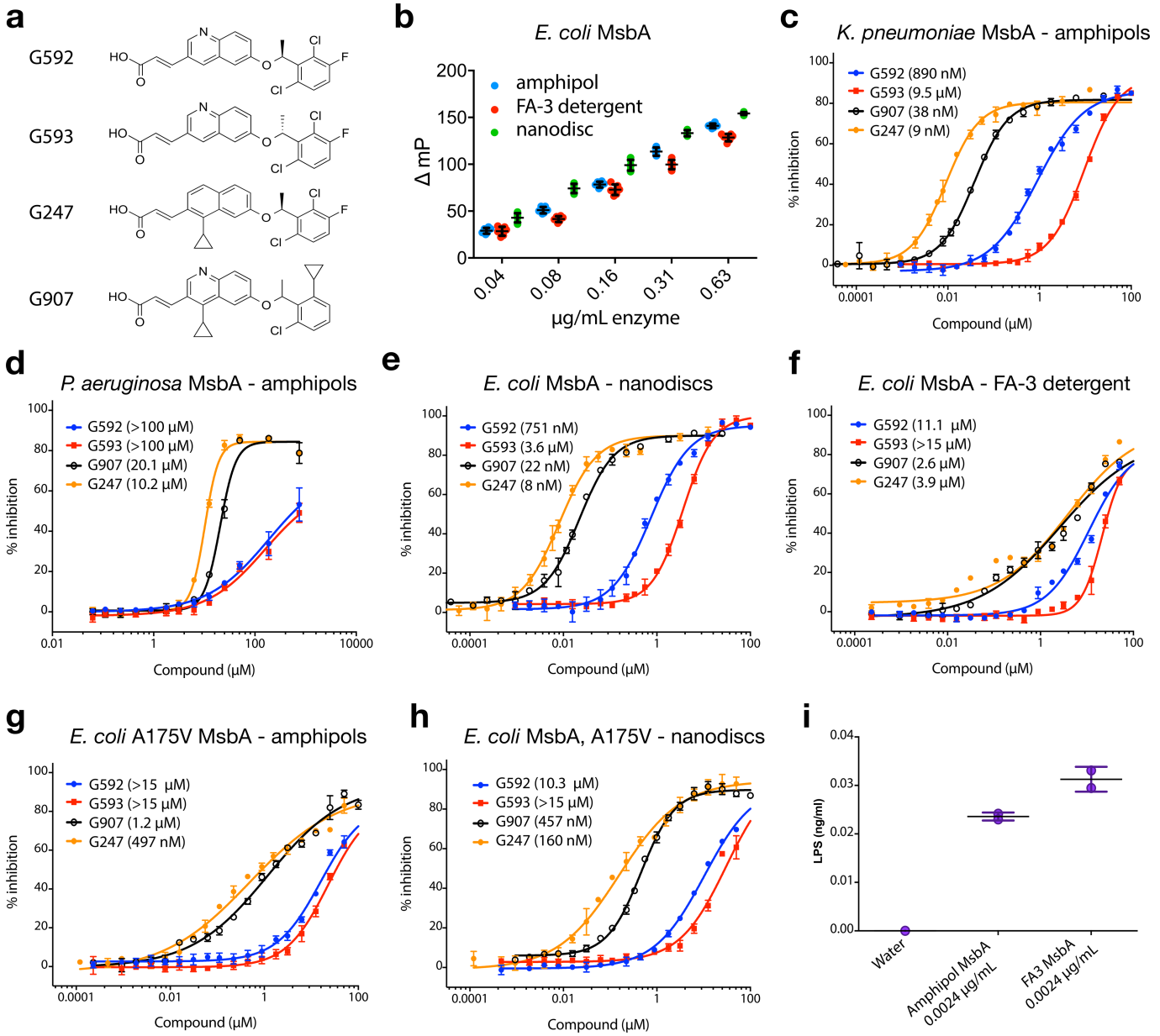
Data availability. The data that support the findings of this study are available from the corresponding authors upon reasonable request. The crystallographic datasets for the G907–LPS–*Ec*MsbA and G092–LPS–*Ec*MsbA complexes that support the findings of this study have been deposited in the Worldwide Protein Data Bank repository (https://wwpdb.org/) under PDB ID accession codes 6BPL and 6BPP, respectively.

26. Zoonens, M. & Popot, J.-L. Amphipols for each season. *J. Membr. Biol.* **247**, 759–796 (2014).
27. Bayburt, T. H. & Sligar, S. G. Membrane protein assembly into nanodiscs. *FEBS Lett.* **584**, 1721–1727 (2010).
28. Diederich, L., Rasmussen, L. J. & Messer, W. New cloning vectors for integration in the lambda attachment site attB of the *Escherichia coli* chromosome. *Plasmid* **28**, 14–24 (1992).
29. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
30. Sampson, B. A., Misra, R. & Benson, S. A. Identification and characterization of a new gene of *Escherichia coli* K-12 involved in outer membrane permeability. *Genetics* **122**, 491–501 (1989).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

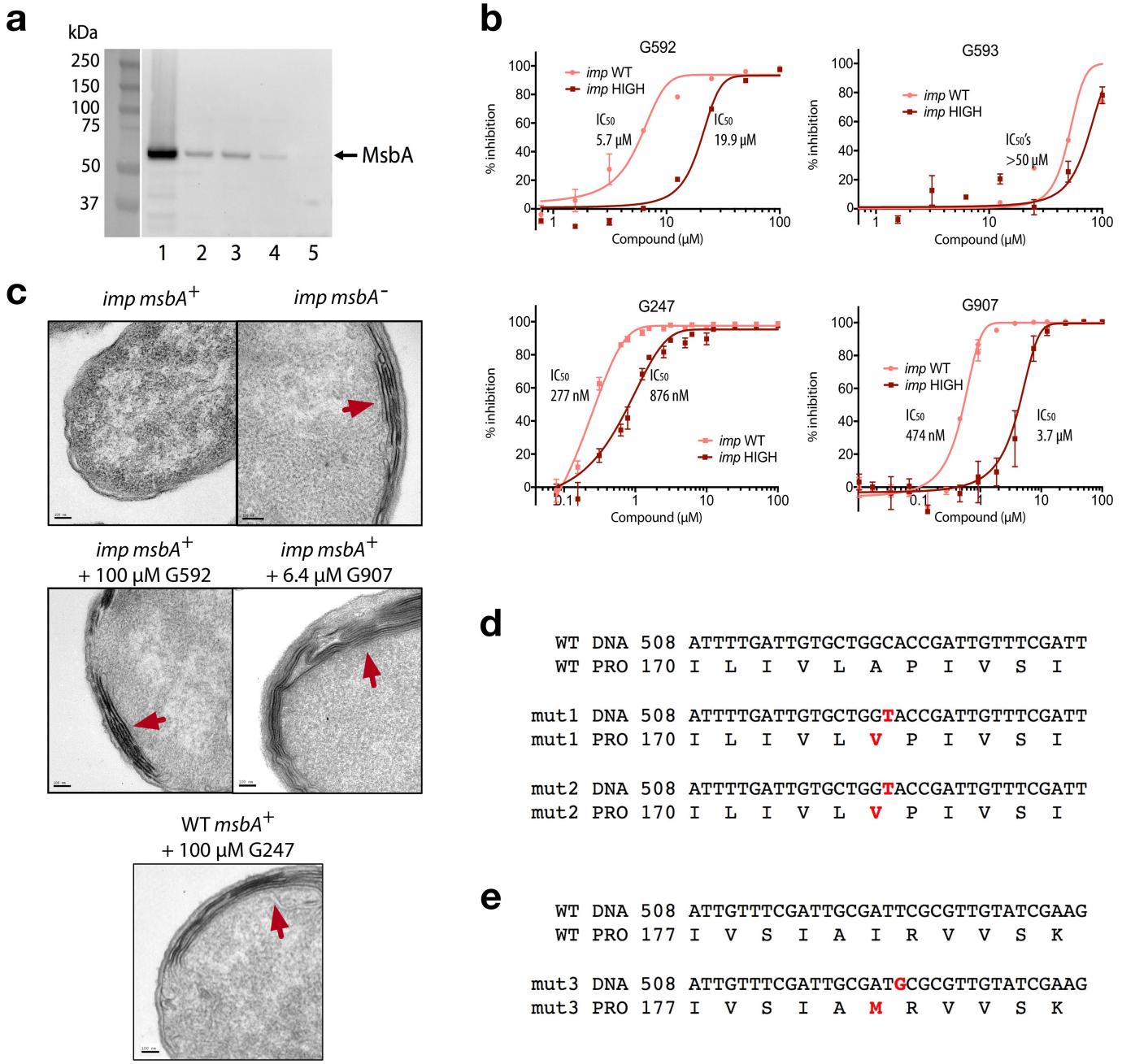
31. Baba, T. et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
32. Wayne, P. A. *Performance Standards for Antimicrobial Susceptibility Testing; Sixteenth Informational Supplement*. CLSI document M100-S16CLSI. (Clinical and Laboratory Standards Institute, Wayne, 2007).
33. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
34. Lawrence, M., Degenhardt, J. & Gentleman, R. VariantTools: Tools for Working with Genetic Variants. R package v.1.20.0 (2016).
35. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
36. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
37. Chaptal, V. et al. Fluorescence detection of heavy atom labeling (FD–HAL): a rapid method for identifying covalently modified cysteine residues by phasing atoms. *J. Struct. Biol.* **171**, 82–87 (2010).
38. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, (213–221 (2010).
39. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
40. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
41. Kleywegt, G. J. Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallogr. D* **52**, 842–857 (1996).
42. The PyMOL Molecular Graphics System v.1.8 (Schrödinger, LLC., 2015).
43. Zhang, Y., Wecksler, A. T., Molina, P., Deperalta, G. & Gross, M. L. Mapping the binding interface of VEGF and a monoclonal antibody Fab-1 fragment with fast photochemical oxidation of proteins (FPOP) and mass spectrometry. *J. Am. Soc. Mass Spectrom.* **28**, 850–858 (2017).
44. Li, J. et al. Mapping the energetic epitope of an antibody/interleukin-23 interaction with hydrogen/deuterium exchange, fast photochemical oxidation of proteins mass spectrometry, and alanine shave mutagenesis. *Anal. Chem.* **89**, 2250–2258 (2017).
45. Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **27**, 221–234 (2013).
46. Schrödinger Release 2017-1: Schrödinger Suite 2017-2 (Schrödinger, LLC, 2017).
47. Schrödinger Release 2017-1: Desmond Molecular Dynamics System (Schrödinger, LLC, 2017).
48. Bowers, K. J. et al. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proc. ACM/IEEE Conference on Supercomputing (SC06)* (ed. Horner-Miller, B.) (ACM, 2006).
49. Shivakumar, D. et al. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field. *J. Chem. Theory Comput.* **6**, 1509–1519 (2010).
50. Guo, Z. et al. Probing the α-helical structural stability of stapled p53 peptides: molecular dynamics simulations and analysis. *Chem. Biol. Drug Des.* **75**, 348–359 (2010).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



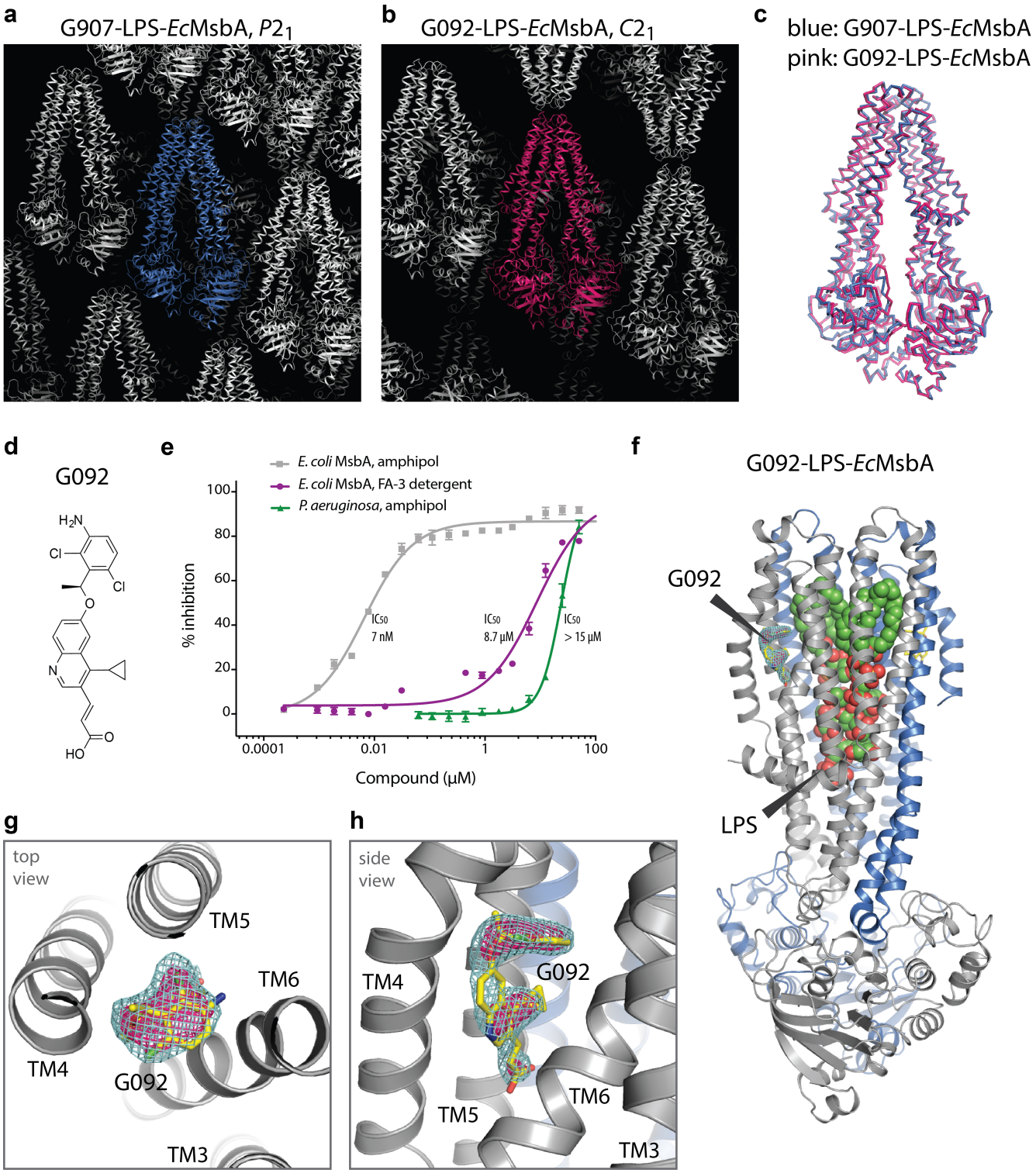
Extended Data Fig. 1 | Evaluation of the biochemical activity of MsbA and inhibitors. a, Quinoline inhibitors described in this study. **b**, Comparison of ATPase activity of increasing concentrations of *E. coli* MsbA in amphipol, nanodisc and FA-3 detergent matrices. **c**, **d**, Dose–response curves of compounds on purified *Klebsiella pneumoniae* (**c**) and *Pseudomonas aeruginosa* (**d**) MsbA reconstituted in amphipols. **e**, **f**, Dose–response curves of compounds on purified *E. coli* MsbA reconstituted in nanodiscs (**e**) or FA-3 detergent (**f**). **g**, **h**, Dose–response curves of compounds on purified *E. coli* A175V mutant MsbA reconstituted in amphipols (**g**) and nanodiscs (**h**). **i**, The LPS content of purified *E. coli* MsbA in the indicated matrices was measured using a chromogenic endotoxin quantitation assay (see Methods section 'Quantitation of co-purifying LPS'). LPS from *E. coli* strain O111:B4 was used to generate a standard curve. Note that for both MsbA samples (0.0024 ng ml^{-1}), the concentration of LPS (more than 0.02 ng ml^{-1}) exceeds the concentration of MsbA, indicating a molar excess of co-purifying LPS. Data are mean ± s.e.m. from five independent experiments (**b**), three independent experiments (**c**–**h**) or two independent experiments (**i**). IC_{50} values (**b**–**h**, in parentheses) were determined by fitting the inhibition dose–response curve with a nonlinear four-parameter inhibition model (see Methods section 'MsbA ATPase assay').

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



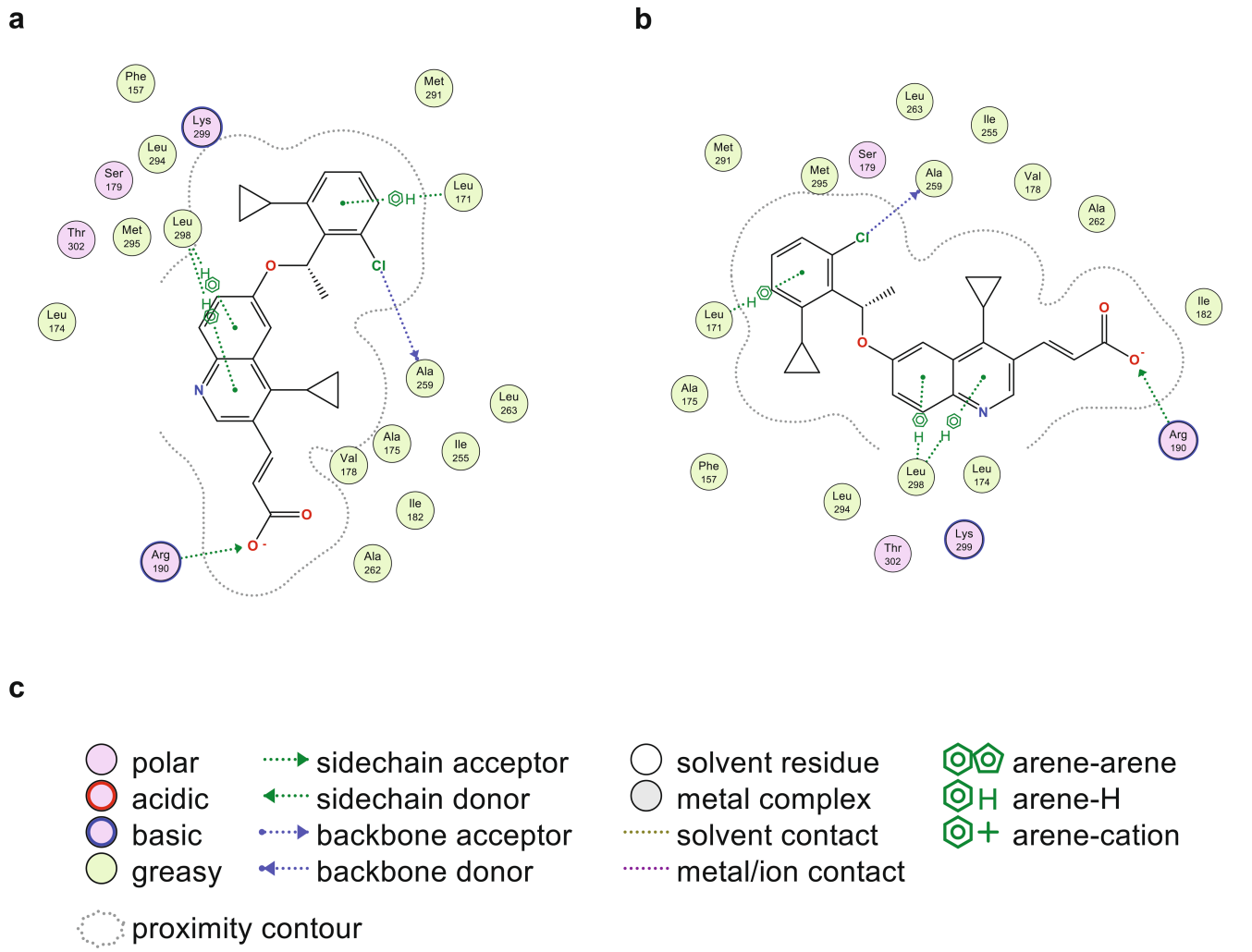
Extended Data Fig. 2 | Evaluation of phenotypic activity of MsbA inhibitors. a, α-MsbA western blot of solubilized extracts from an *E. coli* CFT073 UPEC *imp* strain overexpressing MsbA from a pBAD24 vector in the presence of 2% arabinose (lane 1), an *E. coli* CFT073 *imp* strain (lane 2), an *E. coli* CFT073 *imp* strain with the arabinose-titratable *msbA* expression system in the presence of 2% arabinose (lane 3), an *E. coli* CFT073 *imp* strain with the arabinose-titratable *msbA* expression system in the presence of 0.01% arabinose (lane 4) and an *E. coli* CFT073 *imp* strain with the arabinose-titratable *msbA* expression system in the presence of 0.002% arabinose (lane 5). Bacteria were grown for 3 h at 37 °C before collection. The blot is representative of $n = 2$ independent experiments. For gel source images, see Supplementary Fig. 1. **b**, Dose–response curves of compounds on *E. coli* CFT073 *imp* strains expressing endogenous (WT) and overexpressed (HIGH, in the presence of 0.2% arabinose) levels of MsbA. Data are mean ± s.e.m. from two independent experiments except for G247 ($n = 4$). IC_{50} values (in parentheses) were determined by fitting the inhibition dose–response curve with a nonlinear four-parameter inhibition model (see Methods section 'MsbA ATPase assay'). **c**, Representative (20 isolated cells from two independent experiments) thin section electron micrographs of CFT073 *lptD(imp4213)* and CFT073 wild-type strains. Treatment of the *lptD(imp4213)* strain with G592 (middle left) or G907 (middle right) and wild-type UPEC with G247 (bottom) leads to membrane stacking (red arrow) and enlargement of cells, similar to cells depleted of MsbA (top right). Scale bars, 0.1 μm. **d**, **e**, Nucleotide and amino acid sequence changes (indicated in red) in two G592-resistant MG1655 Δ*tolC* clones (mut1 and mut2, **d**) and a G247-resistant wild-type *E. coli* clone (mut3, **e**, data not shown). For both panels, positions are relative to the start of *msbA* ORF. The frequency of resistance for G247 on wild-type *E. coli* was 3.9×10^{-9} (data not shown).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



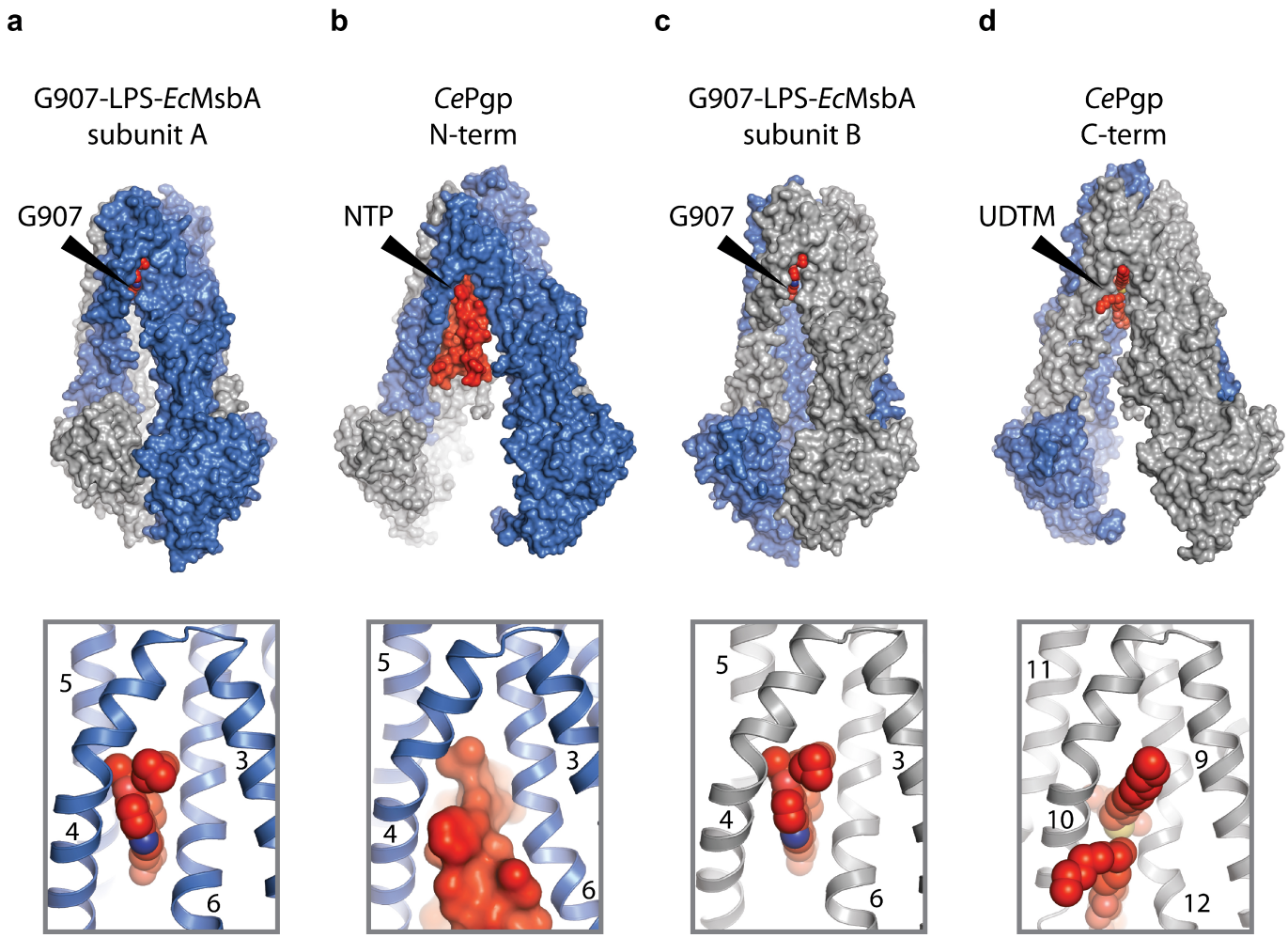
Extended Data Fig. 3 | Comparison of the G907–LPS–*Ec*MsbA and G092–LPS–*Ec*MsbA complexes. a, **b**, Views of the crystal packing environments in the G907–LPS–*Ec*MsbA and G092–LPS–*Ec*MsbA complexes show distinct crystal lattice contacts. **c**, The superposition of G907–LPS–*Ec*MsbA and G092–LPS–*Ec*MsbA complexes indicates a high degree of structural similarity, reflected in a Cα r.m.s.d. of 0.8 Å over all modelled residues. **d**, Chemical structure of G092. **e**, Dose–response curves of G092 on purified *E. coli* MsbA in FA-3 detergent or reconstituted into amphipols and on purified *P. aeruginosa* MsbA reconstituted into amphipols. Data are mean ± s.e.m. from three independent experiments. IC_{50} values (indicated) were determined by fitting the inhibition dose–response curve with a nonlinear four-parameter inhibition model (see Methods). **f**, Overall 2.92-Å crystal structure of the G092–LPS–*Ec*MsbA complex shown from a side-view. LPS bound within the inner vestibule is shown in green sphere representation and G092 is shown in thin yellow stick representation. $F_o - F_c$ difference electron density maps (calculated before the addition of G092 into the model and refinement) are contoured at 2σ (cyan) and 3σ (pink). **g**, $F_o - F_c$ difference electron density maps of G092–LPS–*Ec*MsbA (calculated before the addition of G092 into the model and refinement) are contoured at 2σ (cyan) and 3σ (pink). The final refined coordinates of G092 are shown for reference. **h**, As in **g**, but a side-view of the G092 binding site.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



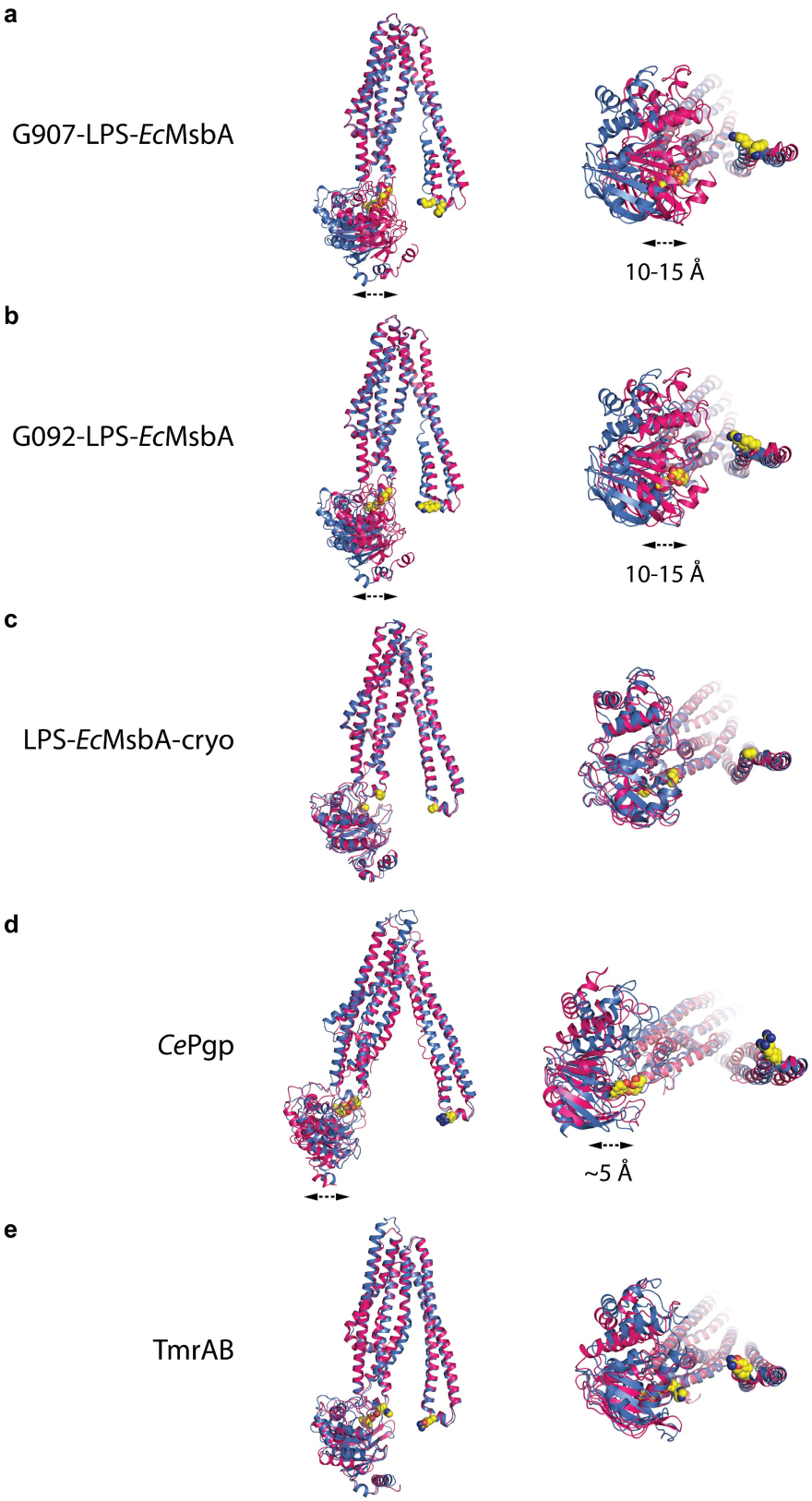
Extended Data Fig. 4 | Molecular interactions of G907 in complex with *Ec*MsbA. a, **b**, The MOE (Molecular Operating Environment; Chemical Computing Group ULC) software was used to highlight the molecular interactions observed between G907 and MsbA, with two similar illustrations of this representation provided. **c**, A legend describing the characteristics of the interacting groups as well as the types of interactions shown in **a** and **b**.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 5 | A conserved small molecule or peptide transmembrane binding pocket in the MsbA and Pgp B-family ABC transporters. a, Top, surface representation of the G907–LPS–*Ec*MsbA crystal structure (subunit A, blue; subunit B, grey) with G907 in red sphere representation. LPS is omitted for clarity. Bottom, expanded view of the G907-binding site. **b**, Top, *Ce*Pgp crystal structure in surface representation (N-terminal half, blue; C-terminal half, grey) with the extended N-terminal peptide (NTP) region of the transporter shown in red surface representation. Bottom, expanded view of the NTP-binding site. **c**, Similar to **a**, G907–LPS–*Ec*MsbA crystal structure in surface representation (subunit A, blue; subunit B, grey) with G907 in red sphere representation, viewed approximately 180° relative to **a** to focus in on subunit B. LPS is omitted for clarity. Bottom, expanded view of the G907-binding site. **d**, Similar to **a**, *Ce*Pgp crystal structure in surface representation (N-terminal half, blue; C-terminal half, grey) in which two modelled detergent molecules of *n*-undecyl-β-D-thiomaltopyranoside (UDTM) are shown in red sphere representation. Bottom, expanded view of the UDTM-binding site.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

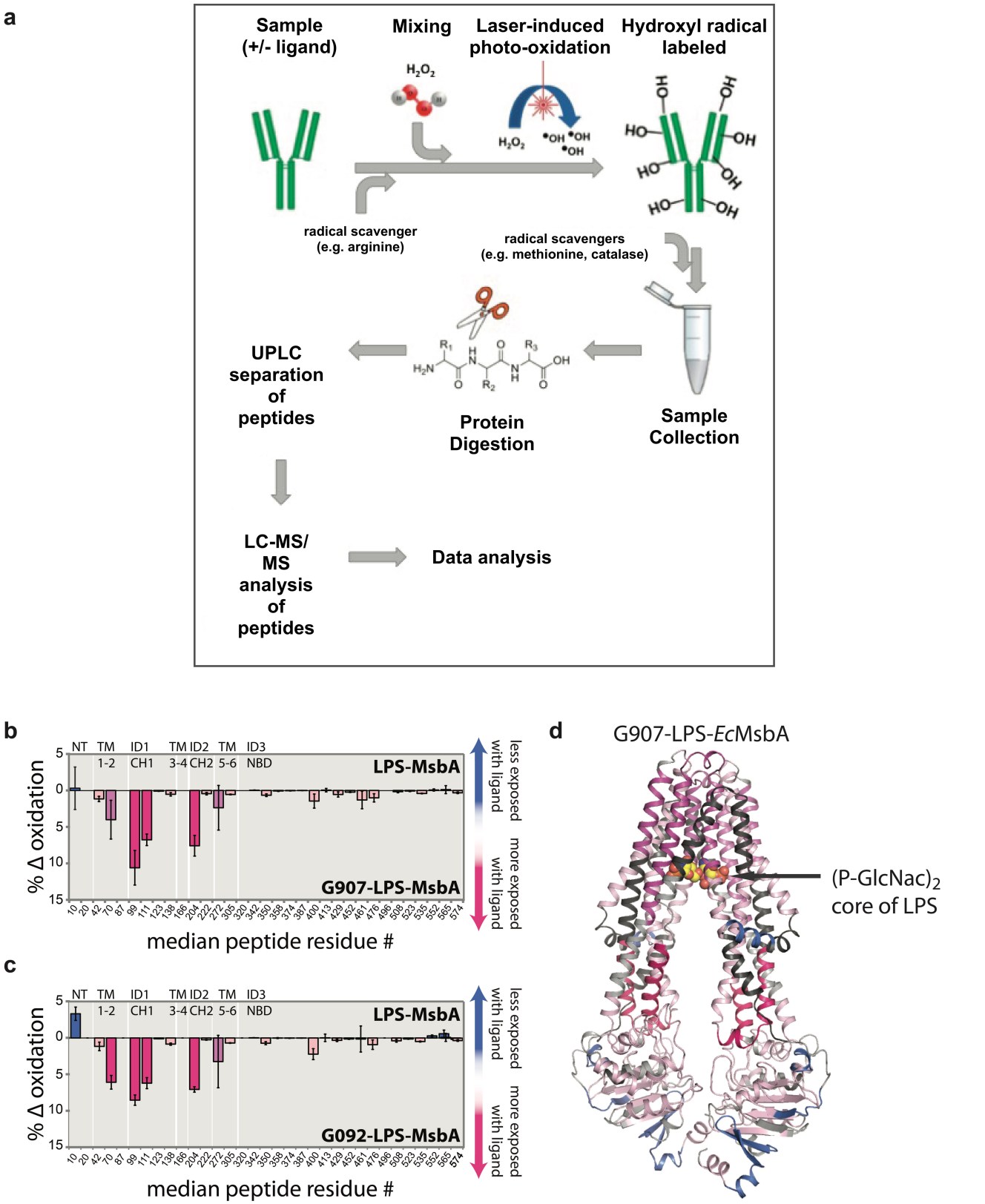


Extended Data Fig. 6 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

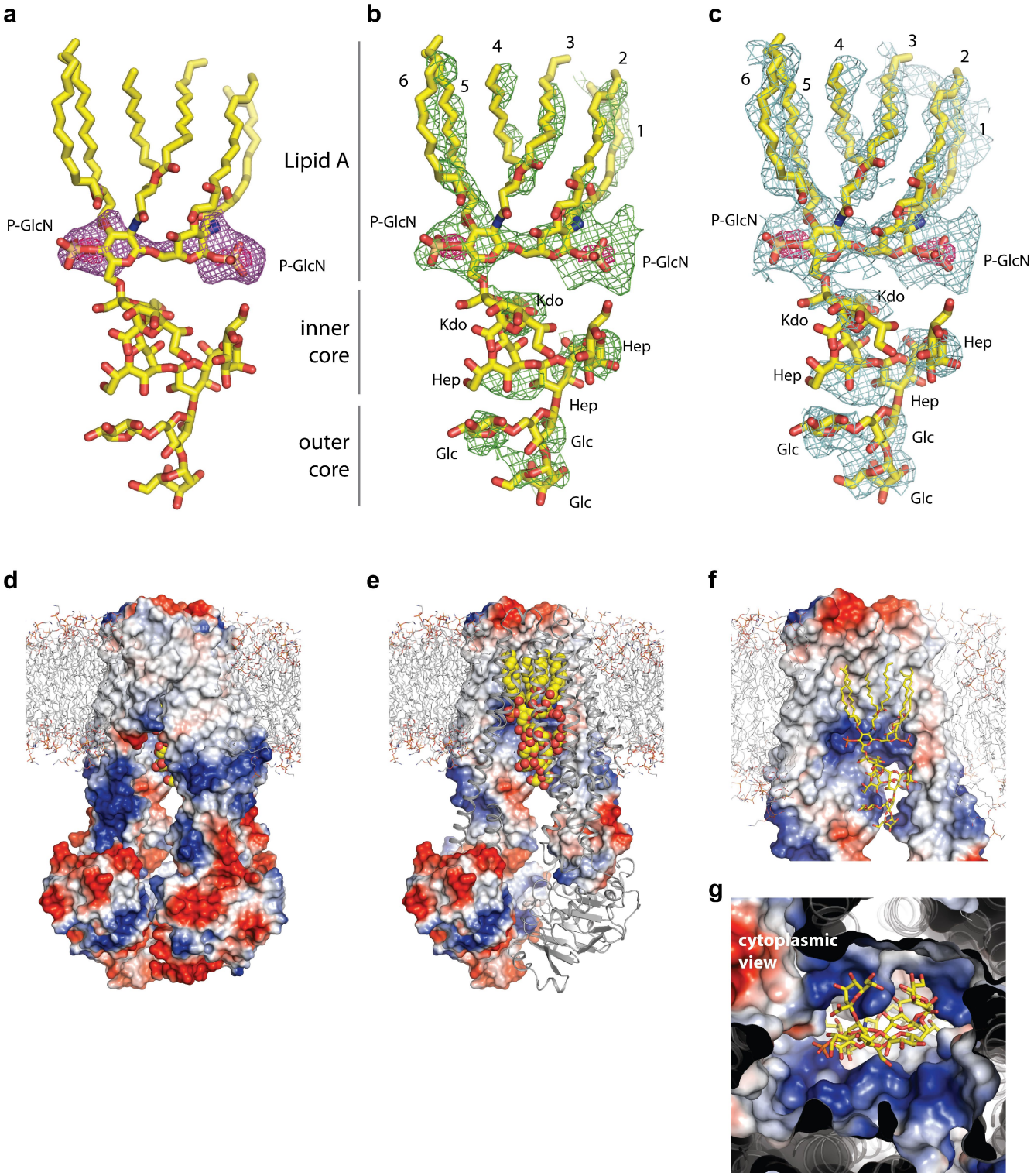
Extended Data Fig. 6 | Unique relative displacement of NBDs in both the G907–LPS–*Ec*MsbA and G092–LPS–*Ec*MsbA complex crystal structures. a–e, Structural superpositions of related B-family ABC transporters are provided in two orthogonal views, in which all structural superpositions were performed by aligning only the transmembrane portion of each subunit B protomer onto that of subunit A. In the case of *Ce*Pgp, the C-terminal transmembrane region was superimposed onto the N-terminal transmembrane region. The structures are as follows: **a**, G907–LPS–*Ec*MsbA; **b**, G092–LPS–*Ec*MsbA; **c**, LPS–*Ec*MsbA–cryo (PDB accession 5TV4); **d**, *Ce*Pgp (PDB accession 4F4C); **e**, *Thermus thermophilus* TmrAB (PDB accession 5MKK). In comparing these structures, it is noteworthy that *Ec*MsbA is a homodimeric transporter, whereas *Ce*Pgp is a concatenated transporter, and TmrAB is a heterodimeric transporter; yet the structural similarity within all 'self-pairs' of transmembrane regions is high. Specifically, when excluding the NBDs from this analysis: **a**, Cα r.m.s.d. 1.26 Å; **b**, Cα r.m.s.d. 0.82 Å; **c**, Cα r.m.s.d. 1.09 Å; **d**, Cα r.m.s.d. 3.18 Å; **e**, Cα r.m.s.d. 1.66 Å. In contrast to the high structural correspondence observed between their superimposed or 'self-paired' transmembrane regions, the G907–LPS–*Ec*MsbA and G092–LPS–*Ec*MsbA complex structures both show a marked translational displacement of one NBD relative to the other, measuring up to 15 Å between equivalent Cα positions. It is notable that the extent and apparent vector of the structural translation observed at the level of the NBDs are unique to the inhibitor-bound *Ec*MsbA complexes studied here. For reference, the three residues that constitute the highly conserved CH1–CH2–NBD coupling interaction network are shown in yellow sphere representation in all structures.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | HRF–MS of LPS–*Ec*MsbA in the presence and absence of G907 and G092. a, A schematic overview (with the target protein illustrated as a cartoon IgG shown in green) of the HRF–MS experiments performed in this study on LPS–*Ec*MsbA in FA-3 detergent conditions in the presence or absence of G907 and G092. LC–MS/MS, liquid chromatography with tandem mass spectrometry; UPLC, ultra-performance liquid chromatography. **b**–**c**, Data are displayed as the differences measured in tryptic peptide oxidation of the ligand-bound condition with G907 present (**b**) or G092 present (**c**) relative to the inhibitor-free LPS–*Ec*MsbA solution sample. The median residue number of the individual tryptic peptides (summarized in Extended Data Table 2 for the G907 experiment) is indicated on the *x*-axis. For both HRF–MS analyses, data are mean ± s.e.m. from three replicate experiments. **d**, A cartoon rendering of the G907–LPS–*Ec*MsbA crystal structure coloured according to the measured differences in tryptic peptide oxidation between the inhibitor-free and G907-bound condition presented in **b**. Regions without significant measured changes are coloured in dark grey and are found mainly within the transmembrane region; regions lacking robust peptide coverage are coloured in light grey. The phospho-*N*-acetylglucosamine moieties of the bound LPS molecule are shown only for reference, the remainder of LPS and G907 are omitted for clarity.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 8 | Electron densities and electrostatic features of the LPS coordination site in the inner vestibule of the G092–LPS–*Ec*MsbA crystal structure. a, $F_o - F_c$ difference electron density maps (calculated before the addition of LPS into the model and refinement) are contoured at 2.5σ (purple) and 5σ (pink). **b**, $F_o - F_c$ difference electron density maps (calculated with LPS omitted from the model) contoured at 3σ (purple) and 5σ (pink). **c**, $F_o - F_c$ difference electron density maps (calculated with LPS omitted from the model) contoured at 1σ (cyan) and 5σ (pink). **a**–**c**, The final refined coordinates of LPS are also shown for reference and the arginine side chains are omitted for clarity. The observed electron density positions the phospho-*N*-acetylglucosamine moieties of bound LPS within the conserved arginine coordination ring. **d**, **e**, Overall electrostatic surface representation of MsbA in the G092–LPS–*Ec*MsbA complex with LPS shown in yellow stick representation and G092 omitted for clarity. For reference only, the position of the bulk membrane bilayer is depicted on the basis of a 10-ns molecular dynamics simulation (see Methods section 'Molecular dynamics'). **f**, Similar to **d**, **e**, expanded view with the closest MsbA subunit removed for clarity and LPS shown in yellow stick representation. **g**, Similar to **f**, expanded view from the cytoplasmic perspective, where both MsbA subunits are displayed.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Data collection, phasing and refinement statistics

	Native G907-LPS-*Ec*MsbA	Native G092-LPS-*Ec*MsbA	Hg-SAD[b] G907-LPS-*Ec*MsbA
Data collection			
Space group	$P2_1$	$C2_1$	$P1$
Cell dimensions			
a, *b*, *c* (Å)	74.7, 91.6, 111.0	256.3, 89.8, 72.5	71.4, 89.2, 109.7
α, *β*, *γ* (°)	90, 89.4, 90	90, 91.3, 90	90.3, 90.1, 89.9
			Hg-Peak
Wavelength	1.000	1.000	1.005
Resolution (Å)[a]	40 – 2.90 (3.00-2.95; 2.95-2.90)	40 – 2.92 (3.02-2.97; 2.97-2.92)	22 – 3.73 (3.8 - 3.73)
R_{merge}	0.097 (0.50; 0.53)	0.098 (0.55; 0.57)	0.03 (0.42)
$I/\sigma(I)$	11.97 (1.09; 0.95)	10.81 (1.17; 1.1)	7.48 (1.31)
$CC_{1/2}$	0.995 (0.624; 0.576)	0.996 (0.553; 0.509)	0.997 (0.513)
Completeness (%)	91.1 (70.4; 61.0)	92.3 (83.3; 78.0)	94.4 (90.4)
Redundancy	4.0 (2.6; 2.4)	3.9 (2.7; 2.6)	1.7 (1.6)
Refinement			
Resolution (Å)	39.9 – 2.91	38.2 – 2.92	
No. reflections	29,686 (1528)	32,865 (2137)	
R_{work} / R_{free}[c]	23.33 / 28.29	25.8 / 27.77	
No. atoms[d]			
subunit A	4469	4389	
subunit B	4358	4124	
G907/G092	62	60	
LPS	233	225	
PE	58	53	
B factors[c]			
subunit A	73.47	61.85	
subunit B	80.00	72.85	
G907	62.41	48.07	
LPS	121.93	116.58	
PE	93.34	80.62	
r.m.s deviations			
Bond lengths (Å)	0.003	0.004	
Bond angles (°)	0.666	0.792	

Single crystal used for each dataset.
[a]Values in parentheses are for highest-resolution shells.
[b]Anomalous signal, 0.1158; SKEW, 0.023; CORR_RMS = 0.738; FOM after SAD phasing/density modification, 0.285/0.73.
[c]5% of reflections used to calculate R_{free}.
[d]Represents all non-hydrogen atoms modelled: native-G907 Ramachandran plot, 95.42% favoured, 4.58% allowed, 0% outlier; native-G092 Ramachandran plot: 95.4% favoured, 4.6% allowed, 0% outlier.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | List of tryptic peptides measured in Fig. 4e HRF–MS experiments

Median peptide residue, as indicated on Figure 4e	Tryptic peptide residues: start - end	Tryptic peptide sequence (median residue indicated in bold and underlined)
10	6-14	DLST**W**QTFR
20	16-24	LWPT**I**APFK
42	25-58	AGLIVAGVALILNAASD**T**FMLSLLKPLLDDGFGK
70	62-78	SVLVWMPL**V**VIGLMILR
87	79-95	GITSYVSS**Y**CISWVSGK
99	96-101	VVM**T**MR
111	104-118	LFGHMMG**M**PVSFFDK
123	119-127	QSTG**T**LLSR
138	128-148	ITYDSEQVAS**S**SSGALITVVR
166	149-183	EGASIIGLFIMMFYYSW**Q**LSIILIVLAPIVSIAIR
204	195-212	NMQNTMGQV**T**TSAEQMLK
222	216-228	EVLIFG**G**QEVETK
272	244-299	MVSASSISDPIIQLIASLALAFVLYAAS**F**PSVMDSLTAGTITVVFSSMIALMRPLK
305	300-310	SLTNV**N**AQFQR
320	311-328	GMAACQTLF**T**ILDSEQEK
342	338-345	ATGD**V**EFR
350	346-354	NVTF**T**YPGR
358	355-360	DVP**A**LR
374	371-377	TVA**L**VGR
387	383-391	STIA**S**LITR
400	392-408	FYDIDEGE**I**LMDGHDLR
413	409-416	EYTL**A**SLR
429	417-441	NQVALVSQNVHL**F**NDTVANNIAYAR
452	448-455	EQIE**E**AAR
461	456-465	MAYAM**D**FINK
476	466-486	MDNGLDTVIG**E**NGVLLSGGQR
496	494-497	AL**L**R
508	498-517	DSPILILDEA**T**SALDTESER
523	518-528	AIQAA**L**DELQK
535	531-538	TSLV**I**AHR
552	545-561	ADEIVVV**E**DGVIVER
565	560-569	GTHND**L**LEHR
574	570-577	GVYA**Q**LHK

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0091-5

Electron magnetic reconnection without ion coupling in Earth's turbulent magnetosheath

T. D. Phan[1]*, J. P. Eastwood[2], M. A. Shay[3], J. F. Drake[4], B. U. Ö. Sonnerup[5], M. Fujimoto[6], P. A. Cassak[7], M. Øieroset[1], J. L. Burch[8], R. B. Torbert[9], A. C. Rager[10,11], J. C. Dorelli[11], D. J. Gershman[11], C. Pollock[12], P. S. Pyakurel[3], C. C. Haggerty[3], Y. Khotyaintsev[13], B. Lavraud[14], Y. Saito[6], M. Oka[1], R. E. Ergun[15], A. Retino[16], O. Le Contel[16], M. R. Argall[9], B. L. Giles[11], T. E. Moore[11], F. D. Wilder[15], R. J. Strangeway[17], C. T. Russell[17], P. A. Lindqvist[18] & W. Magnes[19]

Magnetic reconnection in current sheets is a magnetic-to-particle energy conversion process that is fundamental to many space and laboratory plasma systems. In the standard model of reconnection, this process occurs in a minuscule electron-scale diffusion region[1,2]. On larger scales, ions couple to the newly reconnected magnetic-field lines and are ejected away from the diffusion region in the form of bi-directional ion jets at the ion Alfvén speed[3–5]. Much of the energy conversion occurs in spatially extended ion exhausts downstream of the diffusion region[6]. In turbulent plasmas, which contain a large number of small-scale current sheets, reconnection has long been suggested to have a major role in the dissipation of turbulent energy at kinetic scales[7–11]. However, evidence for reconnection plasma jetting in small-scale turbulent plasmas has so far been lacking. Here we report observations made in Earth's turbulent magnetosheath region (downstream of the bow shock) of an electron-scale current sheet in which diverging bi-directional super-ion-Alfvénic electron jets, parallel electric fields and enhanced magnetic-to-particle energy conversion were detected. Contrary to the standard model of reconnection, the thin reconnecting current sheet was not embedded in a wider ion-scale current layer and no ion jets were detected. Observations of this and other similar, but unidirectional, electron jet events without signatures of ion reconnection reveal a form of reconnection that can drive turbulent energy transfer and dissipation in electron-scale current sheets without ion coupling.

The magnetosheath region, which is located downstream of Earth's bow shock and consists of shocked solar wind, is highly turbulent. This region often contains hundreds of small-scale current sheets in which magnetic reconnection could potentially occur[9,10,12] (Fig. 1c). Many of these sheets are thin (ion inertial length scales or smaller), typically convecting past an observing spacecraft in a few seconds or less. If standard reconnection (Fig. 1a) were to operate in turbulent current sheets, then the ion jets in the extended exhausts should be the easiest reconnection signature to detect. Although electric-field and magnetic-field structures consistent with standard reconnection have been reported[9,10], in situ plasma measurements of the jets were not obtained for these thin current sheets because the data resolution using instruments on previous spacecraft (typically a few seconds per velocity measurement) was not sufficient to determine their presence or absence.

The four-spacecraft Magnetospheric Multiscale (MMS) mission, launched in 2015 and designed to reveal the kinetic physics of reconnection in near-Earth space, is flying in an electron-skin-depth-scale (about 7–10 km) tetrahedral formation. It measures three-dimensional electron and ion distributions at up to 7.5-ms and 37.5-ms resolution[13], respectively, which are 400 times and 80 times better resolved than previously available data. MMS observations of current sheets in the turbulent magnetosheath have revealed thin current sheets[14], fast electron flows[15,16] and electron heating[12,16]. These characteristics are similar to those observed in the electron-diffusion region of standard reconnection in large-scale current sheets at the outer edge of the magnetosphere (the magnetopause)[2,17] and in the laminar magnetosheath (which originate in the solar wind)[18,19]. However, ion jets, which should occur over a larger scale and therefore be more easily observed if standard reconnection is present in the current sheets in the turbulent magnetosheath, remain elusive. This raises the question of whether fast electron flows in these current sheets are produced by some process(es) besides reconnection.

Here we report the serendipitous simultaneous multi-spacecraft detection of oppositely directed super-ion-Alfvénic electron jets, parallel electric fields and magnetic-to-particle energy conversion in an electron-scale current sheet in the magnetosheath, providing direct evidence for reconnection without ion-scale coupling in turbulence.

In Fig. 2a–c we show the large-scale context of the MMS observations in the subsolar magnetosheath region on 2016 December 9 at 8:58–9:43 UT, with large fluctuations in both the magnitude of the magnetic field (Fig. 2a) and its components (Fig. 2b). Figure 2d–g reveals these fluctuations to be sharp changes in the magnetic field associated with large spikes in current density $\boldsymbol{j}$, many with $|\boldsymbol{j}| > 2\,\mu\mathrm{A\ m^{-2}}$ and comparable to the peak current densities observed in the electron-diffusion region at Earth's magnetopause[2,17,20]. Such large current densities across magnetic-field variations of a few tens of nanotesla imply current-sheet widths of a few tens of kilometres or less, less than the ion inertial length d_i of about 50 km in this interval. Closer inspection of the variations in current density and magnetic field throughout the interval in Fig. 2 reveals a range of current-sheet thicknesses: many, but not all, are of sub-ion scales.

To distinguish the regular fast electron flows associated with any thin current sheet from electron jets due to reconnection, data should be examined in a current-sheet ('LMN') coordinate system: the current-sheet normal points along $\boldsymbol{N}$, $\boldsymbol{L}$ is along the anti-parallel magnetic-field direction and $\boldsymbol{M} = \boldsymbol{N} \times \boldsymbol{L}$ is in the out-of-plane ('X-line') direction (Fig. 1a). In such a frame, the main current is in the M direction, while the bi-directional reconnection outflows are in the $\pm L$ directions (Fig. 1b). Definitive evidence for reconnection would be the simultaneous detection of oppositely directed plasma outflow jets by two spacecraft located on opposite sides of the X-line[4].

Such an event was captured at around 09:03:54 UT, when $|\boldsymbol{j}|$ reached approximately $3\,\mu\mathrm{A\ m^{-2}}$ (red arrow in Fig. 2g). In Fig. 3 we show this current sheet in detail, which had a magnetic shear of 14° (the guide

[1]Space Sciences Laboratory, University of California, Berkeley, CA, USA. [2]The Blackett Laboratory, Imperial College London, London, UK. [3]University of Delaware, Newark, DE, USA. [4]University of Maryland, College Park, MD, USA. [5]Dartmouth College, Hanover, NH, USA. [6]ISAS/JAXA, Sagamihara, Japan. [7]West Virginia University, Morgantown, WV, USA. [8]Southwest Research Institute, San Antonio, TX, USA. [9]University of New Hampshire, Durham, NH, USA. [10]Catholic University of America, Washington, DC, USA. [11]NASA Goddard Space Flight Center, Greenbelt, MD, USA. [12]Denali Scientific, Healy, AK, USA. [13]Swedish Institute of Space Physics, Uppsala, Sweden. [14]Institut de Recherche en Astrophysique et Planétologie, Université de Toulouse, Toulouse, France. [15]University of Colorado LASP, Boulder, CO, USA. [16]CNRS/Ecole Polytechnique, Paris, France. [17]University of California, Los Angeles, Los Angeles, CA, USA. [18]Royal Institute of Technology, Stockholm, Sweden. [19]Space Research Institute, Austrian Academy of Sciences, Graz, Austria. *e-mail: phan@ssl.berkeley.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

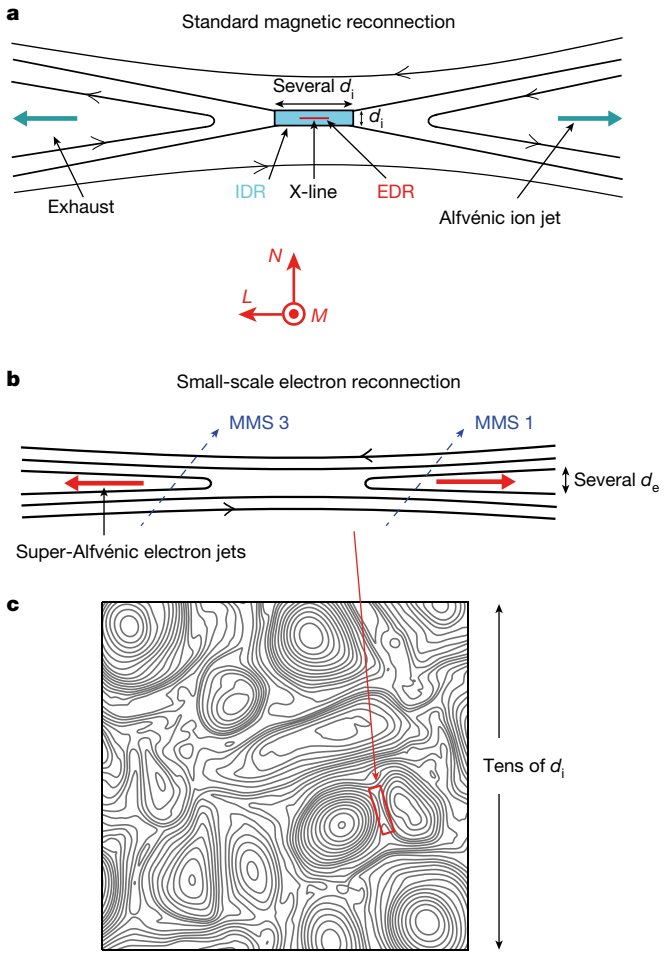


Fig. 1 | Schematics contrasting standard magnetic reconnection in large-scale current sheets and electron-only reconnection in small-scale turbulence. The reconnection configurations in **a** and **b** are displayed in the current-sheet (LMN) coordinate system. **a**, In standard reconnection, the magnetic topology changes in the small electron-diffusion region (EDR) around the X-line, but most of the magnetic-to-particle energy conversion happens in the extended exhausts, with bi-directional ion jetting and heating. The width of the electron-diffusion region (along N) is of the order of the electron skin depth d_e, whereas its length along $\pm L$ could be up to an ion inertial length, $d_i = 43d_e$. The electron-diffusion region is embedded in an ion-diffusion region (IDR), whereas the magnetohydrodynamic-scale reconnection exhaust can extend thousands of d_i (along L) away from the X-line[5]. **b**, Schematic of reconnection in an electron-scale current sheet involving only electrons, with no ion coupling. The entire current sheet is essentially the electron-diffusion region, having a single (electron) scale with embedded bi-directional super-ion-Alfvénic jets. MMS 1 and MMS 3 trajectories through the current sheet relative to the X-line are overlaid (dashed arrows), deduced on the basis of the directions of the electron jets observed on 2016 December 9 at around 09:03:54 UT (and shown in Fig. 3). MMS 1 and MMS 3 were on opposite sides of the X-line detecting bi-directional electron jets. The slanted spacecraft trajectories take into account the likely motion (in the spacecraft frame) of the X-line due to the presence of an external electron flow along $+L$ of about 150 km s^{-1}. **c**, Schematic showing the formation of multiple small (d_i-scale) magnetic structures and thin (d_e-scale) current sheets at their interfaces in turbulent plasmas, informed by turbulence simulations[8,12].

field was $B_M \approx 40$ nT, compared to an anti-parallel field of $|B_L| \approx 5$ nT). Four-spacecraft timing analysis finds the thickness of the current sheet to be only 4 km (or 4 electron skin depths, d_e), determined from the crossing duration of 45 ms (between the vertical dashed lines in Fig. 3) and the convection speed of $V_N = 95$ km s^{-1} of the current sheet.

Inside this electron-scale current sheet, both MMS 3 (left) and MMS 1 (right) observed fast out-of-plane electron flows ($V_{eM} \approx 900$ km s^{-1}; Fig. 3c, m) that produced the main current j_M (Fig. 3d, n) and the associated reversal of B_L (Fig. 3a, k). The speed V_{eM} is comparable to the inflow electron-Alfvén speed of $V_{AeL} = 1{,}000$ km s^{-1} based on $B_L \approx 5$ nT and an electron number density of 20 per cubic centimetre.

Coincident with the intense current layers, MMS 3 and MMS 1 simultaneously observed oppositely directed electron jets in the outflow (L) direction, with $\Delta V_{eL} \approx +250$ km s^{-1} at MMS 3 (Fig. 3c) and $\Delta V_{eL} \approx -450$ km s^{-1} at MMS 1 (Fig. 3m), relative to an external electron flow in the L direction of $V_{eL} \approx +150$ km s^{-1}. The speeds of these electron outflow jets were roughly 10–18 times the asymptotic ion-Alfvén speed (based on B_L) of $V_{AiL} \approx 25$ km s^{-1}. As expected for a reconnection geometry with inflow from both sides, the changes in B_L for MMS 1 are correlated with those in V_{eL} in the first part of the field change and anti-correlated in the second half, whereas for MMS 3 the reverse holds. An exception to this behaviour is that MMS 3, but not MMS 1, observed an electron jet with $\Delta V_{eL} \approx -300$ km s^{-1} at the right edge of the current sheet that is opposite to the main ΔV_{eL} jet (Fig. 3c). Simulations of standard reconnection with a strong guide field have shown such ΔV_{eL} edge jet[21]. The lack of a reversed jet at the edge of the current sheet at MMS 1 is currently not understood.

The measurements of oppositely directed electron outflow jets at MMS 1 and MMS 3 are further supported by the higher-resolution (0.125 ms) measurements of the L component of the field-line velocity $\boldsymbol{E} \times \boldsymbol{B}/B^2$ (where $\boldsymbol{E}$ is the electric field and $\boldsymbol{B}$ is the magnetic field), which was negative at MMS 1 and positive at MMS 3 (except for a negative dip at the right edge, similar to that in V_{eL}) (Fig. 3q, g). These $(\boldsymbol{E} \times \boldsymbol{B}/B^2)_L$ outflows were predominantly perpendicular to the magnetic field, owing to the large B_M (Fig. 3a, k); E_N (Fig. 3e, o), which is opposite at the two spacecraft, together with the dominant B_M, drives the outflows. Crucially, MMS 3 was located 7.1 km in the $+L$ direction relative to MMS 1 so that the observations are consistent with diverging jets from a reconnection X-line located between the two spacecraft as they pass through the reconnecting current sheet. There was no evidence for ion jets (ΔV_{iL}) at the ion-Alfvén speed (Fig. 3b, l) within the thin current sheet. That ion jets are absent is not surprising, because the thickness of the current sheet was only $0.09d_i$ (or 0.08 ion gyroradius) and the observations were made within $7d_e$ of the X-line.

What is surprising, however, is that the electron-scale reconnecting current sheet was not embedded inside a much larger ion-scale current sheet, as would be expected (and observed) in standard reconnection[1,18–20,22] (Fig. 1a). The absence of an outer ion-scale current sheet can be seen in Fig. 3a, k (see also Extended Data Fig. 1), which shows B_L reaching its asymptotic values immediately outside the thin current sheet.

Both spacecraft detected well defined parallel electric fields (Fig. 3f, p), which implies that the 'electron frozen-in condition' ($\boldsymbol{E}' = \boldsymbol{E} + \boldsymbol{V}_e \times \boldsymbol{B} = \boldsymbol{0}$) was violated. Furthermore, $\boldsymbol{j} \cdot \boldsymbol{E}'$ was positive (Fig. 3j, t) and dominated by $j_{\parallel}E_{\parallel}$, indicating non-ideal magnetic-to-particle energy conversion[23] characteristic of the electron-diffusion region. However, unlike for standard reconnection, where most of the magnetic energy is converted into ion jetting and heating, here, half of the available magnetic energy per particle (half of 6 eV) in the inflow regions ($m_e(V_{AeL})^2$, where m_e is the electron mass) goes into kinetic energy associated with ΔV_{eM} and ΔV_{eL}, which are 90% and 45% of V_{AeL}, respectively. The remaining half (3 eV), if converted entirely into electron heating, would lead to an increase in electron temperature of (3 eV) × $(\gamma-1)/\gamma \approx 1$ eV in the reconnecting current sheet[24], where $\gamma = 5/3$ is the ratio of specific heats. Such a small temperature increase would not be discernible in the data (Fig. 3i, s).

Within the 21-min burst data intervals shown in Fig. 2a, there were 34 other isolated current sheets with $|\boldsymbol{j}| > 2\,\mu$A m^{-2}, which implies sub-ion-scale current-sheet widths. Surprisingly, the majority of these current sheets had low magnetic shears (that is, strong guide fields): 23 of the 34 events had magnetic shear of less than 45°. All 34 sheets had a fast out-of-plane electron velocity V_{eM} consistent with the large current density j_M, but only 16 displayed clear super-ion-Alfvénic V_{eL} jets that could be related to reconnection. In each of these cases, all four spacecraft detected V_{eL} pointing in the same direction and were

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

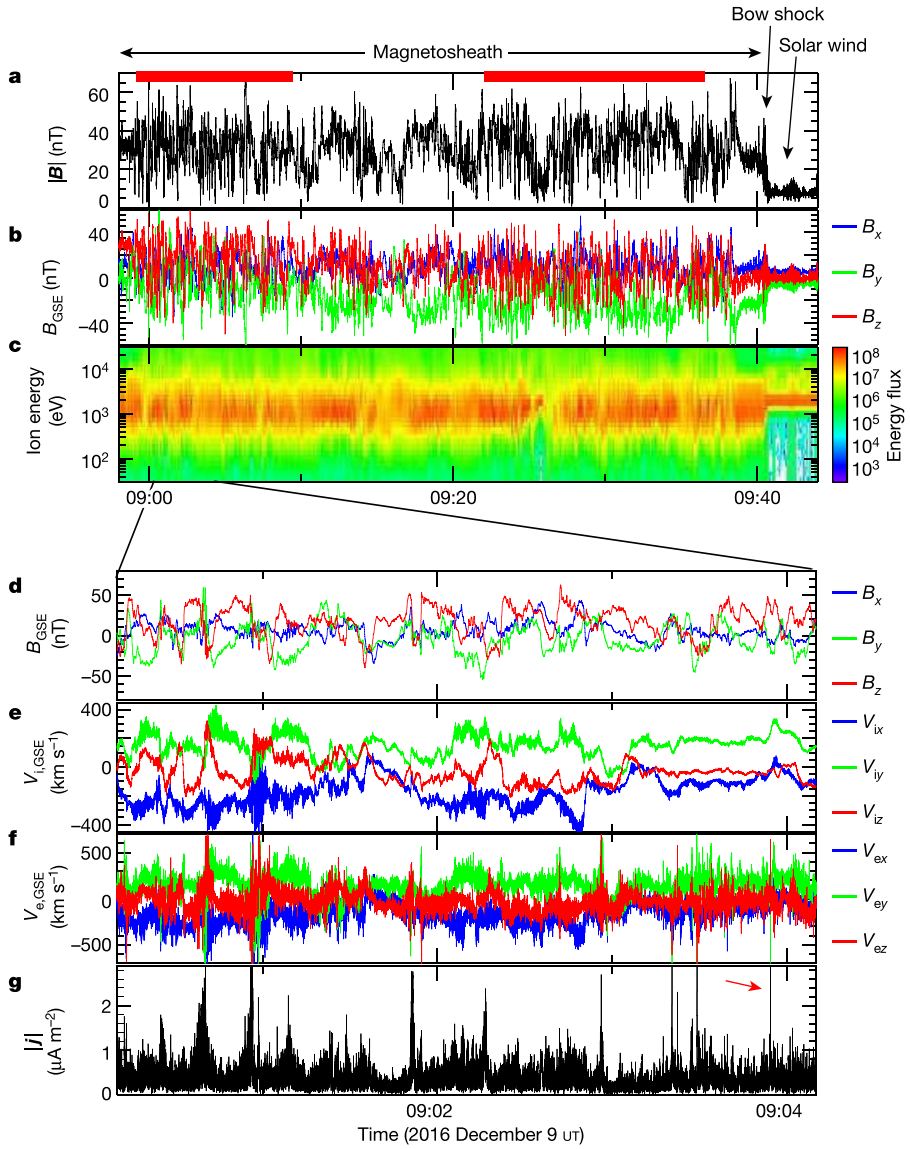


Fig. 2 | Overview of MMS 1 observations of turbulent current sheets in Earth's subsolar magnetosheath region downstream of a quasi-parallel shock. The observations reveal the presence of large spikes in current density (more than 2 μA m^{-2}), which imply sub-ion-scale current sheets. The data are displayed in the geocentric solar ecliptic (GSE) coordinates. **a**, **b**, The magnitude (**a**; $|B|$) and components (**b**; B_x, B_y and B_z) of the magnetic field. c, The ion energy–time spectrogram of differential energy flux (colour scale, in units of eV s^{-1} cm^{-2} sr^{-1} eV^{-1}). **d**–**g**, Zoomed-in (4-min) interval showing the magnetic field (**d**; B_x, B_y and B_z), ion velocity (**e**; V_{ix}, V_{iy} and V_{iz}), electron velocity (**f**; V_{ex}, V_{ey} and V_{ez}) and current density (**g**; $|\boldsymbol{j}| = \boldsymbol{j} = eN_e(V_i - V_e)$, where e is the elementary charge and N_e is the electron number density) computed from plasma measurements. Throughout the interval in **a**–**c**, the angle between the magnetic field of the solar wind and the Sun–Earth line was less than 30° and the subsolar bow shock was quasi-parallel. The spike in current density at around 09:03:54 UT (indicated by the red arrow in **g**) is the bi-directional electron jet event, shown in detail in Fig. 3. Red horizontal bars in **a** denote burst data intervals totalling 21 min, selected by the MMS scientist on duty because they contained a large number of high-amplitude magnetic-field fluctuations (with $\Delta B/B \approx 0.5$), suggestive of current sheets that could be prone to reconnection.

therefore embedded in the same jet. The scarcity of unambiguous reconnection events with divergent jets (which implies that the X-line is located between the spacecraft) is probably due to the small (about 7 km or $7d_e$) separations between the spacecraft.

We found no evidence for reconnection ion jetting associated with any of the electron outflow jet events, or in any other (including ion-scale-thick) current sheets in the 21-min interval (see examples in Extended Data Fig. 2). This finding is in stark contrast to standard models of reconnection in which the ion exhaust jets should be easier to detect than the electron-diffusion region because they extend large distances from the X-line.

The absence of ion reconnection signatures suggests that, in these turbulent magnetosheath plasmas, there is insufficient space and/or time for the ions to couple to the magnetic structures. This could occur not only because the widths of the current sheets are of electron scales, but also if the overall dimensions of the current sheets are limited because ion coupling requires some minimum lengths along the exhaust[25] (L) and X-line[26] (M) directions. A hybrid simulation study of resistive reconnection with no guide field[25] suggests that ions become decoupled when the length of the current sheet in the L direction (Fig. 1c) falls below about $10d_i$. If such an ion decoupling scale also exists in strong-guide-field collisionless reconnection, although potentially at a smaller scale than $10d_i$, then it could account for our observed lack of ion coupling in reconnection in the magnetosheath, in which the coherence scales of magnetic structures have been reported[27] to be of the order of d_i.

The experimental discovery of electron-only reconnection reveals that reconnection operates differently in current sheets with small

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

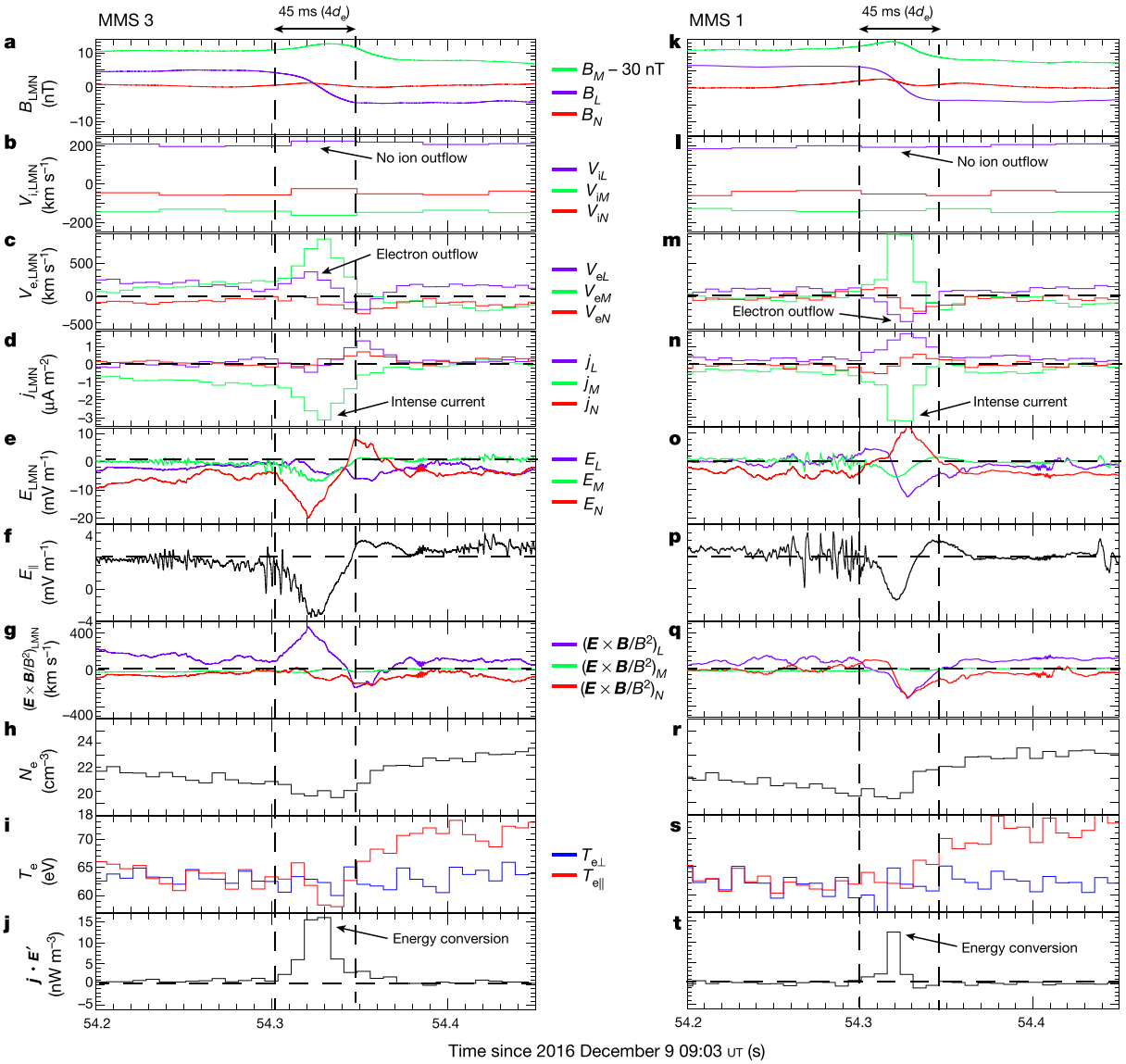


Fig. 3 | Simultaneous MMS 1 and MMS 3 detections of oppositely directed super-ion-Alfvénic electron jets, parallel electric fields and enhanced magnetic-to-electron energy conversion in an electron-scale current sheet. The data for both spacecraft (MMS 3, **a**–**j**; MMS 1, **k**–**t**) are shown in a common current-sheet (LMN) coordinate system, determined for the MMS 3 crossing of the current sheet at 09:03:54.270–09:03:54.365 UT, with $\boldsymbol{L} = (-0.091, 0.87, 0.49)_{GSE}$, $\boldsymbol{M} = (-0.25, -0.49, 0.83)_{GSE}$ and $\boldsymbol{N} = (0.96, -0.05, 0.27)_{GSE}$. The vertical dashed lines mark the left and right edges of the current sheet. **a**, **k**, Magnetic field B_{LMN} at 8,196 samples per second (from merged fluxgate and search-coil magnetometer measurements[28]), with B_M shifted by −30 nT. **b**, **c**, **l**, **m**, Ion (**b**, **l**; $V_{i,LMN}$) and electron (**c**, **m**; $V_{e,LMN}$) velocity. The 7.5-ms electron and 37.5-ms ion data products were generated by separating the individual energy sweeps that were used to form the nominal burst-mode distribution functions. These data maintain sufficient angular coverage to recover accurate plasma moments at four times the nominal temporal resolution[13]. **d**, **n**, Current density (j_{LMN}) from plasma measurements. **e**, **o**, Electric field[28] in the spacecraft frame (E_{LMN}) at 8,196 samples per second. **f**, **p**, Electric-field component parallel to the magnetic field ($E_{||}$). **g**, **q**, $(\boldsymbol{E} \times \boldsymbol{B}/B^2)_{LMN}$ velocity. **h**, Electron density (N_e). **i**, Electron temperature (T_e). **j**, **t**, $\boldsymbol{j} \cdot (\boldsymbol{E} + \boldsymbol{V}_e \times \boldsymbol{B}) = \boldsymbol{j} \cdot \boldsymbol{E}'$. The LMN coordinate system was determined using a hybrid minimum-variance method, which often works best in low-magnetic-shear current sheets[29]. The current-sheet normal direction N was determined from $\boldsymbol{B}_1 \times \boldsymbol{B}_2/|\boldsymbol{B}_1 \times \boldsymbol{B}_2|$, where $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are the fields at the two edges of the current sheet. We define $\boldsymbol{M} = \boldsymbol{L}' \times \boldsymbol{N}$, where $\boldsymbol{L}'$ is the direction of maximum variance of the magnetic field[30]; $\boldsymbol{L} = \boldsymbol{N} \times \boldsymbol{M}$. MMS 3 was located at $L = +7.1$ km, $M = +3.3$ km and $N = +1.6$ km relative to MMS 1. Data from all four spacecraft are shown in Extended Data Fig. 3.

overall dimensions. Our finding supports the long-held idea that reconnection has a role in dissipating the energy associated with plasma turbulence in space and astrophysical systems, although the scale for dissipation by reconnection would be at the electron scale instead of the ion scale. To assess the importance of reconnection in dissipating turbulence energy in small systems quantitatively, the basic properties of electron-only reconnection (such as the rate, duration and onset conditions of reconnection) will need to be investigated theoretically and observationally. These properties could differ substantially from those known from the standard model of reconnection.

Data availability

The entire MMS dataset is publicly available at https://lasp.colorado.edu/mms/sdc/public/.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0091-5.

Received: 7 December 2017; Accepted: 30 January 2018;
Published online 9 May 2018.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

1. Vasyliunas, V. M. Theoretical models of magnetic merging. *Rev. Geophys.* **13**, 303–336 (1975).
2. Burch, J. L. et al. Electron-scale measurements of magnetic reconnection in space. *Science* **352**, aaf2939 (2016).
3. Paschmann, G. et al. Plasma acceleration at the Earth's magnetopause: evidence for reconnection. *Nature* **282**, 243–246 (1979).
4. Phan, T. D. et al. Extended magnetic reconnection at the Earth's magnetopause from detection of bi-directional jets. *Nature* **404**, 848–850 (2000).
5. Gosling, J. T., Skoug, R. M., McComas, D. J. & Smith, C. W. Direct evidence for magnetic reconnection in the solar wind near 1 AU. *J. Geophys. Res.* **110**, A01107 (2005).
6. Petschek, H. E. Magnetic field annihilation. In Proc. AAS-NASA Symposium on the Physics of Solar Flares (ed. Hess, W. N.) 425–439 (NASA, 1964)
7. Matthaeus, W. H. & Lamkin, S. L. Turbulent magnetic reconnection. *Phys. Fluids* **29**, 2513–2534 (1986).
8. Servidio, S., Matthaeus, W. H., Shay, M. A., Cassak, P. A. & Dmitruk, P. Magnetic reconnection in two-dimensional magnetohydrodynamic turbulence. *Phys. Rev. Lett.* **102**, 115003 (2009).
9. Retinò, A. et al. In situ evidence of magnetic reconnection in turbulent plasma. *Nat. Phys.* **3**, 235–238 (2007).
10. Sundqvist, D., Retinò, A., Vaivads, A. & Bale, S. D. Dissipation in turbulent plasma due to reconnection in thin current sheets. *Phys. Rev. Lett.* **99**, 025004 (2007).
11. Haggerty, C. C. et al. Exploring the statistics of magnetic reconnection X-points in kinetic particle-in-cell turbulence. *Phys. Plasmas* **24**, 102308 (2017).
12. Chasapis, A. et al. Electron heating at kinetic scales in magnetosheath turbulence. *Astrophys. J.* **836**, 247 (2017).
13. Rager, A. C. et al. Electron crescent distributions as a manifestation of diamagnetic drift in an electron scale current sheet: magnetospheric multiscale observations using new 7.5 ms fast plasma investigation moments. *Geophys. Res. Lett.* **45**, 578–584 (2018).
14. Eriksson, E. et al. Strong current sheet at a magnetosheath jet: kinetic structure and electron acceleration. *J. Geophys. Res.* **121**, 9608–9618 (2016).
15. Yordanova, E. et al. Electron scale structures and magnetic reconnection signatures in the turbulent magnetosheath. *Geophys. Res. Lett.* **43**, 5969–5978 (2016).
16. Vörös, Z. et al. MMS observations of magnetic reconnection in the turbulent magnetosheath. *J. Geophys. Res.* **122**, 11442–11467 (2017).
17. Chen, L. J. et al. Electron energization and mixing observed by MMS in the vicinity of an electron diffusion region during magnetopause reconnection. *Geophys. Res. Lett.* **43**, 6036–6043 (2016).
18. Phan, T. D., Drake, J. F., Shay, M. A., Mozer, F. S. & Eastwood, J. P. Evidence for an elongated (60 ion skin depths) electron diffusion region during fast magnetic reconnection. *Phys. Rev. Lett.* **99**, 255002 (2007).
19. Wilder, F. D. et al. Multipoint measurements of the electron jet of symmetric magnetic reconnection with a moderate guide field. *Phys. Res. Lett.* **118**, 265101 (2017).
20. Burch, J. L. & Phan, T. D. Magnetic reconnection at the dayside magnetopause: advances with MMS. *Geophys. Res. Lett.* **43**, 8327–8338 (2016).
21. Pritchett, P. L. Geospace environment modeling magnetic reconnection challenge: simulations with a full particle electromagnetic code. *J. Geophys. Res.* **106**, 3783–3798 (2001).
22. Shay, M. A. et al. Structure of the dissipation region during collisionless magnetic reconnection. *J. Geophys. Res.* **103**, 9165–9176 (1998).
23. Zenitani, S., Hesse, M., Klimas, A. & Kuznetsova, M. New measure of the dissipation region in collisionless magnetic reconnection. *Phys. Rev. Lett.* **106**, 195003 (2011).
24. Shay, M. A. et al. Electron heating during magnetic reconnection: a simulation scaling study. *Phys. Plasmas* **21**, 122902 (2014).
25. Mandt, M. E. et al. Transition to whistler mediated magnetic reconnection. *Geophys. Res. Lett.* **21**, 73–76 (1994).
26. Meyer, J. C. *Structure of the Diffusion Region in Three Dimensional Magnetic Reconnection*. PhD thesis, Univ. Delaware (2015).
27. He, J. S. et al. Two-dimensional correlation functions for density and magnetic field fluctuations in magnetosheath turbulence measured by the Cluster spacecraft. *J. Geophys. Res.* **116**, A06207 (2011).
28. Torbert, R. B. et al. The FIELDS instrument suite on MMS: scientific objectives, measurements, and data products. *Space Sci. Rev.* **199**, 105–135 (2016).
29. Gosling, J. T. & Phan, T. D. Magnetic reconnection in the solar wind at current sheets associated with extremely small field shear angles. *Astrophys. J.* **763**, L39 (2013).
30. Sonnerup, B. U. Ö. & Cahill, L. J. Jr. Magnetopause structure and attitude from Explorer 12 observations. *J. Geophys. Res.* **72**, 171–183 (1967).

Acknowledgements We are grateful for the dedicated efforts of the MMS team. We thank J. Gosling, who inspired us to search for unconventional reconnection in space. This work was supported by NASA contract number NNG04EB99C at SwRI, which funded work at most of the co-authors' institutions in the United States. The work at U. C. Berkeley was supported by NASA grants 80NSSC18K0157 and NNX08A083G. UK involvement at Imperial College was supported by STFC (UK) grant ST/N000692/1. The French involvement (SCM instruments) on MMS is supported by CNES, CNRS-INSIS and CNRS-INSU.

Reviewer information *Nature* thanks G. Paschmann and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.D.P. carried out the data analysis, interpretation and manuscript preparation. J.P.E., M.A.S., J.F.D., B.U.Ö.S., M.F., P.A.C., M.Ø., P.S.P., C.C.H., M.O., and A.R. contributed to the data interpretation and manuscript preparation. J.L.B. led the successful design and operation of the MMS mission and contributed to the interpretation of the data. A.C.R., J.C.D., D.J.G., C.P., B.L., B.L.G., T.E.M., and Y.S. contributed to the development and operation of and the interpretation of data from the Fast Plasma instruments. R.B.T., R.E.E., Y.K., M.R.A., F.D.W. and P.A.L. contributed to the development and operation of and the interpretation of data from the electric-field experiments. O.L.C. contributed to the development and operation of the search-coil magnetometers. C.T.R., R.J.S., and W.M. contributed to the development and operation of the fluxgate magnetometers.

Competing interests The authors declare no competing interests.

Additional information

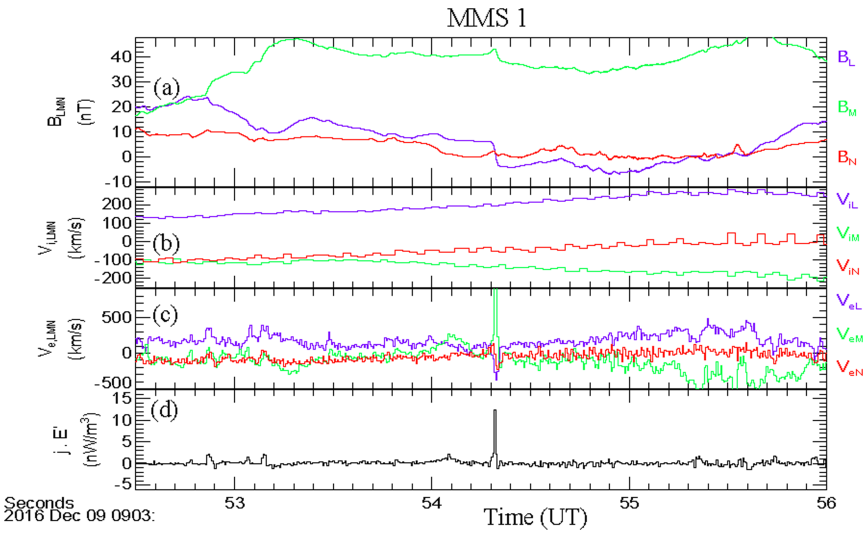
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0091-5.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to T.D.P.

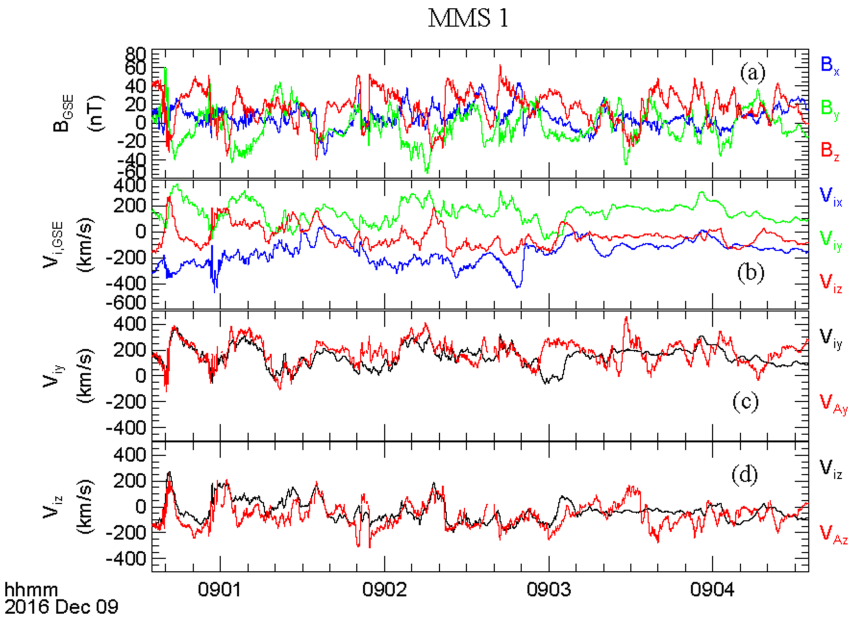
Publisher's note: Springer Nature remains neutral with regard tojurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



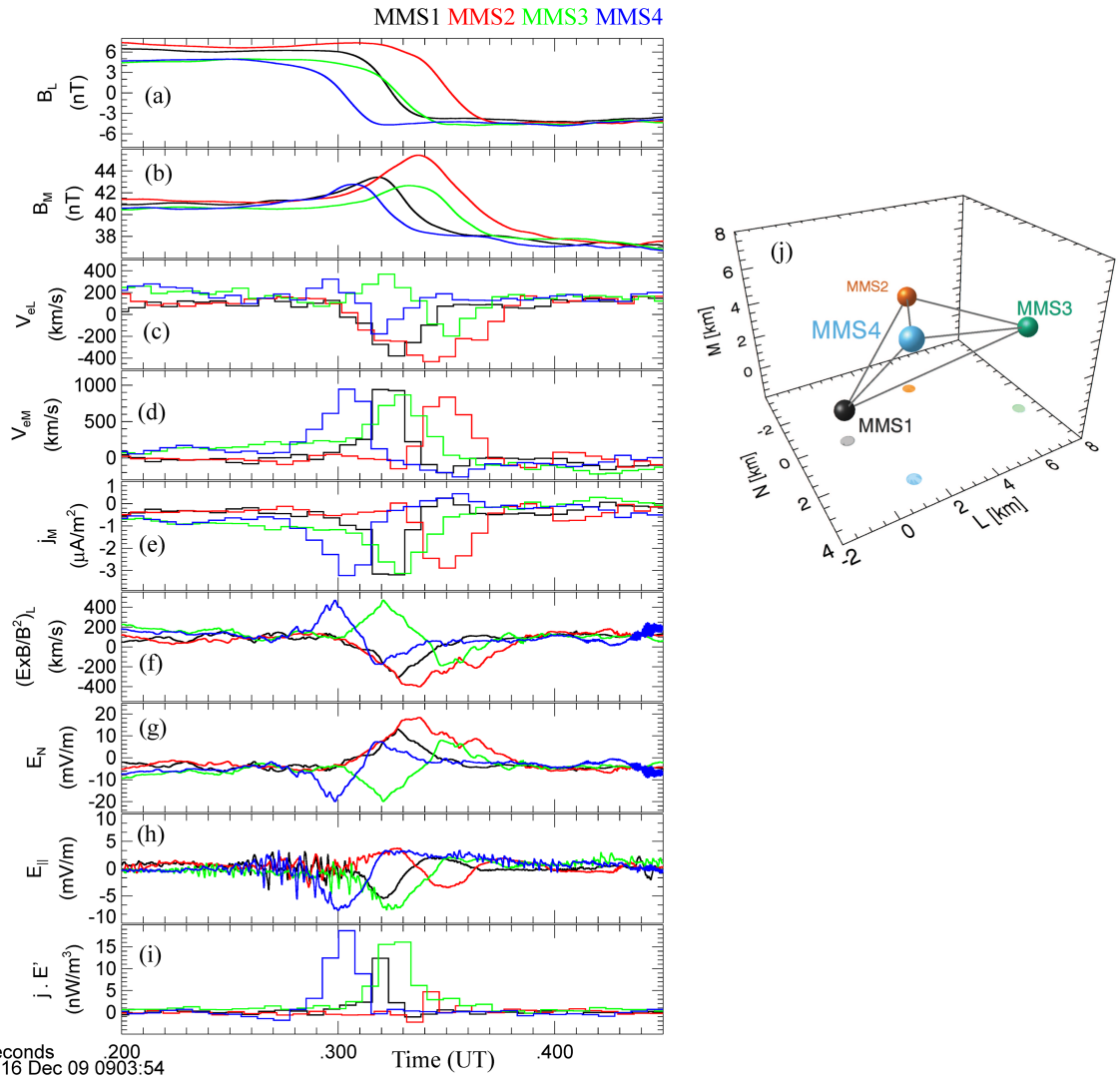
Extended Data Fig. 1 | Large-scale context of the thin current sheet, illustrating the fact that the electron-scale current sheet was a stand-alone current sheet, not embedded inside an ion-scale one. Data are shown in the LMN coordinates determined for the thin current sheet and used in Fig. 3. **a**, Magnetic field. **b**, Ion velocity. **c**, Electron velocity. **d**, $\boldsymbol{j}\cdot(\boldsymbol{E}+\boldsymbol{V}_e\times\boldsymbol{B})=\boldsymbol{j}\cdot\boldsymbol{E}'$. The thin reconnecting current sheet stands out in this interval, with nothing else approaching its current density or its value of $\boldsymbol{j}\cdot(\boldsymbol{E}+\boldsymbol{V}_e\times\boldsymbol{B})$. The absence of an ion-scale current sheet enveloping the electron-scale current sheet is indicated by the fact that $|B_L|$ reaches essentially its asymptotic values immediately outside the thin current sheet.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 2 | Absence of reconnection ion jetting. The data are in GSE coordinates. **a**, Magnetic field. **b**, Ion velocity. **c**, **d**, y (**c**) and z (**d**) components of the ion velocity ($\boldsymbol{V}_i$) and Alfvén velocity ($\boldsymbol{V}_A$). $\boldsymbol{V}_A$ is relative to the reference velocity, density and magnetic-field values at the left edge of the data interval: $\boldsymbol{V}_A = \boldsymbol{B}_{ref}(1-\alpha_{ref})^{1/2}(\mu_0\rho_{ref})^{-1/2} - \boldsymbol{B}(1-\alpha)^{1/2}(\mu_0\rho)^{-1/2}$, where μ_0 is the vacuum permeability, $\alpha = (p_{||}-p_{\perp})\mu_0/B^2$ is the pressure anisotropy factor and ρ is the mass density of the plasma[3]. The expected speeds of the ion reconnection jets embedded inside many of the large-magnetic-shear current sheets are in the range of 100–200 km s^{-1}, based on B_L variations of around 20−40 nT (**a**). If present, such jets are readily recognized by back-to-back opposite correlations between ion velocity and magnetic-field variation at the two edges of the current sheet, which indicate pairs of rotational discontinuities emanating from the X-line[5]. These signatures are not seen here. What we instead find in the data is either no correlation between components of $\boldsymbol{V}_i$ and $\boldsymbol{B}$, or a single correlation (or anti-correlation), indicative of Alfvénic structures[16] rather than reconnection jetting.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 3 | Four-spacecraft observations of the reconnecting current sheet. A common current-sheet LMN coordinate system (same as in Fig. 3) was used for consistency, justified by the fact that the LMN coordinates at individual spacecraft differ from each other by less than 4°. **a**–**d**, L (**a**, **c**) and M (**b**, **d**) components of the magnetic field (**a**, **b**) and of the electron velocity (**c**, **d**). **e**, M component of the current density. **f**, L component of the $\boldsymbol{E}\times\boldsymbol{B}/B^2$ velocity. **g**, N component of the electric field. **h**, Electric-field component parallel to the magnetic field. **i**, $\boldsymbol{j}\cdot(\boldsymbol{E}+\boldsymbol{V}_e\times\boldsymbol{B})=\boldsymbol{j}\cdot\boldsymbol{E}'$. **j**, Spacecraft locations relative to MMS 1, in kilometres (roughly d_e). The B_L profiles (**a**) show that MMS 1 and MMS 3 crossed the current sheet at essentially the same time, preceded by MMS 4 and followed by MMS 2. The fact that MMS 4 exited the current sheet before MMS 2 entered it places an upper limit on the thickness of the current sheet, which is the 4.5-km separation distance between the two spacecraft along $\boldsymbol{N}$ (**j**). This is consistent with the current-sheet width of 4 km determined from the motion and crossing duration of the current sheet. Inside the current sheet, MMS 4 detected a positive $(\boldsymbol{E}\times\boldsymbol{B}/B^2)_L$ (except at the right edge), similarly to MMS 3, whereas MMS 2 detected a negative $(\boldsymbol{E}\times\boldsymbol{B}/B^2)_L$, similarly to MMS 1. This indicates that there was a pair of spacecraft on each side of the X-line. All four spacecraft detected a predominantly negative $E_{||}$. The parameter $\boldsymbol{j}\cdot(\boldsymbol{E}+\boldsymbol{V}_e\times\boldsymbol{B})$ was consistently positive at all four spacecraft throughout the current sheet, with the amplitude being lowest at MMS 2. MMS 2 also detected the largest guide-field (B_M) compression, fastest ΔV_{eL} and $(\boldsymbol{E}\times\boldsymbol{B}/B^2)_L$ jets, slowest ΔV_{eM} and weakest $E_{||}$, which together may suggest that MMS 2 was furthest away from the X-line.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0096-0

Precision measurement of the weak charge of the proton

The Jefferson Lab Q_{weak} Collaboration*

Large experimental programmes in the fields of nuclear and particle physics search for evidence of physics beyond that explained by current theories. The observation of the Higgs boson completed the set of particles predicted by the standard model, which currently provides the best description of fundamental particles and forces. However, this theory's limitations include a failure to predict fundamental parameters, such as the mass of the Higgs boson, and the inability to account for dark matter and energy, gravity, and the matter–antimatter asymmetry in the Universe, among other phenomena. These limitations have inspired searches for physics beyond the standard model in the post-Higgs era through the direct production of additional particles at high-energy accelerators, which have so far been unsuccessful. Examples include searches for supersymmetric particles, which connect bosons (integer-spin particles) with fermions (half-integer-spin particles), and for leptoquarks, which mix the fundamental quarks with leptons. Alternatively, indirect searches using precise measurements of well predicted standard-model observables allow highly targeted alternative tests for physics beyond the standard model because they can reach mass and energy scales beyond those directly accessible by today's high-energy accelerators. Such an indirect search aims to determine the weak charge of the proton, which defines the strength of the proton's interaction with other particles via the well known neutral electroweak force. Because parity symmetry (invariance under the spatial inversion $(x, y, z) \rightarrow (-x, -y, -z)$) is violated only in the weak interaction, it provides a tool with which to isolate the weak interaction and thus to measure the proton's weak charge[1]. Here we report the value 0.0719 ± 0.0045, where the uncertainty is one standard deviation, derived from our measured parity-violating asymmetry in the scattering of polarized electrons on protons, which is -226.5 ± 9.3 parts per billion (the uncertainty is one standard deviation). Our value for the proton's weak charge is in excellent agreement with the standard model[2] and sets multi-teraelectronvolt-scale constraints on any semi-leptonic parity-violating physics not described within the standard model. Our results show that precision parity-violating measurements enable searches for physics beyond the standard model that can compete with direct searches at high-energy accelerators and, together with astronomical observations, can provide fertile approaches to probing higher mass scales.

In the electroweak standard model, elastic scattering is mediated by the exchange of neutral currents (virtual photons and Z^0 bosons) between fundamental particles. A particle's weak charge, Q_w, is analogous to—but distinct from—its electric charge, q; the former quantifies the vector coupling of the Z^0 boson to the particle, whereas the latter quantifies the vector coupling of the photon to the particle. The proton's weak charge Q_w^p is defined[1] as the sum of the weak vector coupling constants C_{1q} of the Z^0 boson to its constituent up (u) and down (d) quarks:

$$Q_w^p = -2(2C_{1u} + C_{1d}) \quad (1)$$

The Z^0 exchange contribution to electron–proton scattering can be isolated via the weak interaction's unique parity-violation signature (see Fig. 1). Interference between electromagnetic and weak scattering amplitudes leads to a parity-violation asymmetry A_{ep} that can be measured with a longitudinally polarized electron beam incident on an unpolarized-proton target:

$$A_{ep} = \frac{\sigma_+ - \sigma_-}{\sigma_+ + \sigma_-} \quad (2)$$

Here, $\sigma_\pm$ represents the cross-section of the helicity-dependent elastic scattering (ep) of polarized electrons (e) on protons (p), integrated over the scattered-electron detector acceptance. Helicity (± 1) indicates the spin direction of the electrons in the beam as either parallel ($+1$) or anti-parallel (-1) to their momenta.

Measuring A_{ep} at a small four-momentum transfer (Q^2) suppresses contributions from the proton's extended structure relative to the proton's weak charge Q_w^p. However, A_{ep} is small at small Q^2 values, making its measurement challenging. In the low-Q^2 limit, the parity-violation asymmetry can be expressed[1] as:

$$A_{ep}/A_0 = Q_w^p + Q^2 B(Q^2, \theta) \quad (3)$$

where $A_0 = -G_F Q^2/(4\pi\alpha\sqrt{2})$, $-Q^2$ is the four-momentum transfer squared, $B(Q^2, \theta)$ represents the proton's extended structure defined in terms of electromagnetic, strange and axial form factors, θ is the (polar) scattering angle of the electron in the laboratory frame with respect to the beam axis, G_F is the Fermi constant and α is the fine-structure constant.

The Q_{weak} experiment[3,4] (see Extended Data Fig. 1) used a beam of longitudinally polarized electrons accelerated to 1.16 GeV at the Thomas Jefferson National Accelerator Facility. Three sequential acceptance-defining lead collimators, matched to an eight-sector azimuthally symmetric toroidal magnet, selected electrons scattered from a liquid-hydrogen target at angles between 5.8° and 11.6°. In each magnet octant, elastically scattered electrons were directed to a quartz detector fronted by lead pre-radiators. Cherenkov light produced by the electromagnetic shower passing through the quartz was converted to a current by photomultiplier tubes at each end of the quartz bars. These currents were integrated for each 1-ms-long helicity state of the beam. For calibration purposes, and to confirm our understanding of the acceptance and backgrounds, drift chambers were periodically inserted upstream and downstream of the magnet to track individual particles during dedicated periods of low-current running.

To achieve a precision of less than 10 parts per billion (p.p.b.), this experiment pushed existing boundaries on many fronts: higher polarized-beam intensity (180 μA), faster beam-helicity reversal (960 s^{-1}), better GeV-scale beam polarimetry[5] precision (±0.6%), higher liquid-hydrogen target[6] luminosity (1.7×10^{39} cm^{-2} s^{-1}) and cooling power (3 kW) and higher total detection rates (7 GHz). Following a brief commissioning period, the experiment was divided into two roughly six-month run periods, between which improvements were made primarily to polarimetry and helicity-related beam-monitoring and control instrumentation. Further details, including the

*A list of participants and their affiliations appears at the end of this Letter.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

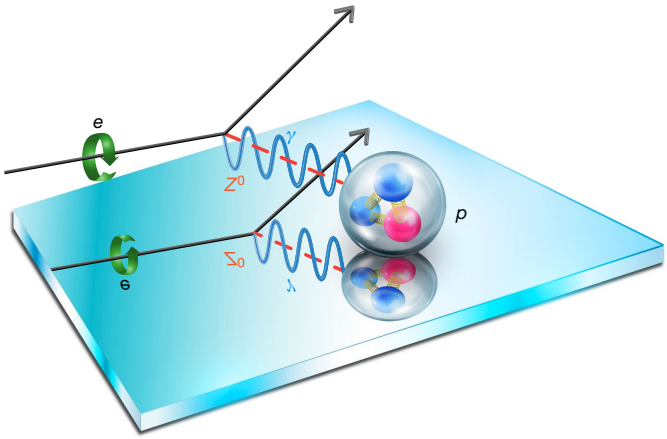


Fig. 1 | Parity-violating electron scattering from the proton. An incoming electron, *e*, with helicity +1 scatters away from the plane of the 'parity-violating mirror'. The image in the parity-violating mirror shows the incoming electron with the opposite helicity, −1; instead of scattering into the plane of the parity-violating mirror (as it would in a real mirror), it scatters out of the plane of the parity-violating mirror. The dominant electromagnetic interaction, mediated by the photon (γ, blue wavy line), conserves parity. The weak interaction, mediated by the neutral Z^0 boson (dashed red line), violates parity. The weak interaction is studied experimentally by exploiting parity violation through reversals of the incident-beam helicity, which mimic the parity-violating mirror 'reflection'.

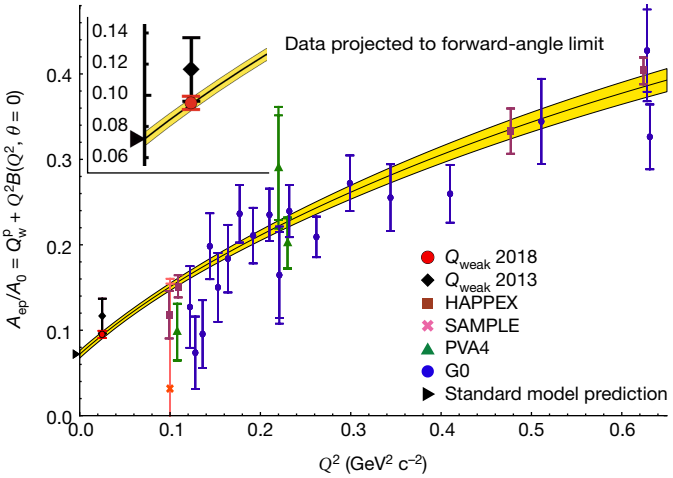


Fig. 2 | The reduced asymmetry $A_{ep}/A_0 = Q_w^p + Q^2B(Q^2, \theta = 0)$ versus Q^2. The global fit is illustrated using *ep* asymmetries from this experiment (Q_{weak} 2018), from the commissioning phase of this experiment[3] (Q_{weak} 2013), as well as from the earlier experiments HAPPEX, SAMPLE, PVA4 and G0 (see Methods), projected to $\theta = 0°$ and reduced by a factor $A_0(Q^2)$ appropriate for each datum. The data shown here include the γZ-box radiative correction and uncertainty. Inner error bars correspond to one standard deviation (s.d.) and include statistical and systematic uncertainties. Outer error bars on the data indicate the additional uncertainty estimated from the forward-angle projection (for some data points, inner and outer error bars coincide). The solid line represents the global fit to the complete PVES database (see Methods), and the yellow band indicates the fit uncertainty (1 s.d.). The arrowhead at $Q^2 = 0$ indicates the standard-model prediction[2], $Q_w^p = 0.0708(3)$, which agrees well with the intercept of the fit ($Q_w^p = 0.0719 \pm 0.0045$). The inset shows a magnification of the region around this experiment's result, at $\langle Q^2 \rangle = 0.0248\ \mathrm{GeV}^2\,c^{-2}$.

backgrounds and corrections associated with each of the two halves of the experiment, are provided in Methods.

The asymmetry measurement results are $A_{ep} = -223.5 \pm 15.0$ (statistical) $\pm$ 10.1 (systematic) p.p.b. in the first half of the experiment, and $A_{ep} = -227.2 \pm 8.3$ (statistical) $\pm$ 5.6 (systematic) p.p.b. in the second half. These values are in excellent agreement with each other and consistent with our previously published commissioning result[3]. Accounting for correlations in some systematic uncertainties between the two measurement periods, the combined result is $A_{ep} = -226.5 \pm 7.3$ (statistical) $\pm$ 5.8 (systematic) p.p.b. The total uncertainty achieved (9.3 p.p.b.) sets a new level of precision for parity-violating electron scattering (PVES) from a nucleus.

The relationship between the measured asymmetries A_{ep} and the proton's weak charge Q_w^p is expressed by equation (3), where the hadronic-structure-dependent term $B(Q^2, \theta)$ grows with the momentum transfer Q^2. Higher-Q^2 data from previous PVES experiments (see online references, Methods) were included in a global fit[3,7,8] to constrain the proton-structure contributions for the short extrapolation from our datum to $Q^2 = 0$ in order to determine Q_w^p, the intercept of equation (3). The average Q^2 of this experiment (0.0248 GeV$^2\,c^{-2}$) is much smaller than that of any other PVES experiments used in this fit, with correspondingly smaller contributions from the proton structure. The superior precision of the Q_{weak} measurement tightly constrains the fit near $Q^2 = 0$, where the connection to Q_w^p can be made.

The parameters of the global fit[3,7,8] to the PVES data are the axial-electron–vector-quark weak-coupling constants C_{1u} and C_{1d}, the strange charge radius ρ_s and strange magnetic moment μ_s (which characterize the strength of the proton's electric and magnetic strange-quark form factors) and the strength of the neutral weak (Z^0 exchange) isovector ($T = 1$) axial form factor $G_A^{Z(T=1)}$. The EM form factors G_E and G_M used in the fit were taken from ref. [9]; uncertainties in this input were accounted for in the result for Q_w^p and in its uncertainty.

The *ep* asymmetries shown in Fig. 2 were corrected[1,3] for the energy-dependent part of the γZ-box weak radiative correction[10–13] and its uncertainty. No other electroweak radiative corrections need to be applied to determine Q_w^p. However, ordinary electromagnetic radiative corrections (bremsstrahlung) were accounted for in the asymmetries used in the fit, including our datum. Details of the fitting procedure, as well as a description of the corrections applied to the asymmetry for this experiment, are described in Methods.

The global fit is shown in Fig. 2 together with the *ep* data, expressed as $A_{ep}(Q^2, \theta = 0)/A_0$. To isolate the Q^2 dependence for this figure, the θ dimension was projected to 0° by subtracting $[A_{calc}(Q^2, \theta) - A_{calc}(Q^2, \theta = 0)]$ from the measured asymmetries $A_{ep}(Q^2, \theta)$, as described in refs [3,8]. Here A_{calc} refers to the asymmetries determined from the global fit. The fit includes all relevant PVES data for the scattering of polarized electrons on protons (*ep*), deuterons (e^2H) and ^{4}He (e^4He); see Methods. The PVES database provides a data-driven (as opposed to a more theoretical) constraint on the nucleon structure uncertainties in the extrapolation to $Q^2 = 0$. We consider this to be the best method to provide our main result (denoted in Table 1 as

Table 1 | Results extracted from the asymmetry measured in the Q_{weak} experiment

Method	Quantity	Value	Error
PVES fit	Q_w^p	0.0719	0.0045
	ρ_s	0.20	0.11
	μ_s	−0.19	0.14
	$G_A^{Z(T=1)}$	−0.64	0.30
PVES fit + APV	Q_w^p	0.0718	0.0044
	Q_w^n	−0.9808	0.0063
	C_{1u}	−0.1874	0.0022
	C_{1d}	0.3389	0.0025
	C_1 correlation	−0.9318	
PVES fit + LQCD	Q_w^p	0.0685	0.0038
Q_{weak} datum only	Q_w^p	0.0706	0.0047
Standard model	Q_w^p	0.0708	0.0003

'PVES fit' refers to a global fit incorporating the Q_{weak} result and the PVES database, as described in Methods. When combined with APV[14,15] (to improve the C_{1d} precision), this method is denoted as 'PVES fit + APV'. If the strange form factors in the global fit (without APV) are constrained to match LQCD calculations[16], we label the result as 'PVES fit + LQCD'. The method labelled 'Q_{weak} datum only' uses the Q_{weak} datum, together with electromagnetic[9], strange[16] and axial[18] form factors from the literature in lieu of the global fit. Uncertainties are 1 s.d.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

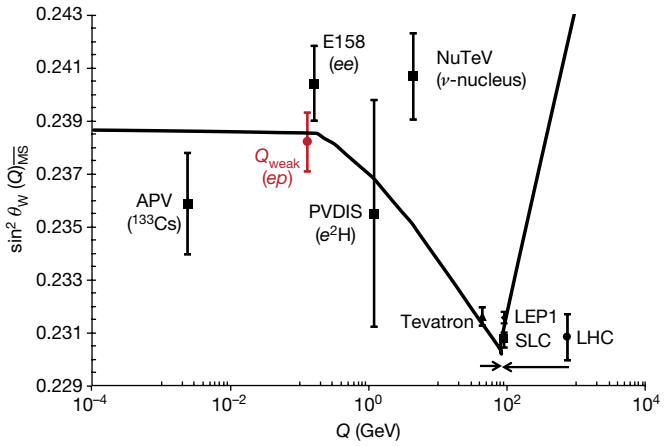


Fig. 3 | Variation of $\sin^2\theta_W$ with energy scale Q. The modified-minimal-subtraction ($\overline{MS}$) scheme is shown as the solid curve[2,19], together with experimental determinations at the Z^0 pole[2] (Tevatron, LEP1, SLC, LHC), from APV on caesium[14,15], Møller scattering (E158)[22], deep inelastic scattering of polarized electrons on deuterons (e^2H; PVDIS)[23] and from neutrino–nucleus scattering (NuTeV)[24]. It has been argued[25], however, that the latter result contains substantial unaccounted-for nuclear physics effects, such as neutron-excess corrections to the quark momenta, charge-symmetry breaking and strange-quark momentum asymmetries. Our new result is plotted in red at the energy scale of the Q_{weak} experiment, $Q = 0.158$ GeV (slightly offset horizontally for clarity). Error bars (1 s.d.) include statistical and systematic uncertainties.

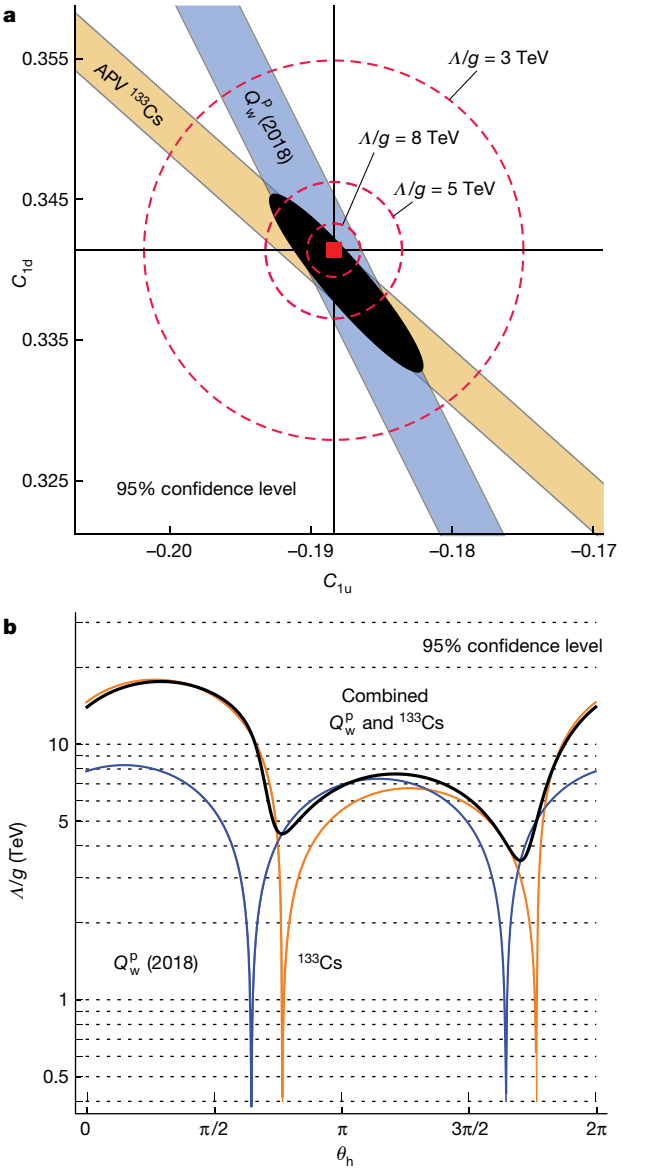


Fig. 4 | Mass and coupling constraints on new physics. a, Constraints, at the 95% confidence level, on the axial-electron–vector-quark weak-coupling constants C_{1u} and C_{1d}, derived from the weak charge determined in this experiment using the global fit method 'PVES fit' (blue band) and the APV result[2,14,15] on ^{133}Cs (gold band). The combined (95% confidence level) constraint is shown by the black ellipse. Contours of the mass reach Λ/g for new physics with coupling g to arbitrary quark-flavour ratios are indicated by dashed circles centred about the standard-model values[2] of C_{1u} and C_{1d}, which are denoted by the red square. **b**, Mass reach Λ/g (95% confidence level) as a function of the quark-flavour mixing angle θ_h for the Q_{weak} 'PVES fit' result (blue curve), for the ^{133}Cs APV[14,15] result[2] (gold curve) and for both results combined (black curve). The two maxima in the blue curve at $\theta_h = \tan^{-1}(n_d/n_u) = \tan^{-1}(1/2) = 26.6°$ and 206.6° correspond to $\Lambda_-/g = 8.4$ TeV and $\Lambda_+/g = 7.4$ TeV in equation (4), respectively.

'PVES fit'), which is $Q_w^p = 0.0719 \pm 0.0045$. Below we discuss the sensitivity of this result to variations in the experimental and theoretical input used to determine it.

Just as the proton's weak charge depends on its u and d quark content (see equation (1)), the weak charge of other nuclear systems depends on their (different) u and d quark content. Because ep, e^2H and e^4He data are included in the global fit, C_{1u} and C_{1d} are reasonably well determined. However, if the very precise atomic-parity violation (APV) result[14,15] on ^{133}Cs is also included in the global fit, C_{1u} and C_{1d} can be determined with greater precision and then used to extract the neutron's weak charge $Q_w^n = -2(C_{1u} + 2C_{1d})$. We note that inclusion or exclusion of the APV result has negligible impact on our result for Q_w^p, which is derived from the intercept of the global fit. The results for C_{1u}, C_{1d}, Q_w^p and Q_w^n obtained by including APV in the PVES global fit, which are listed in Table 1 as 'PVES fit + APV', are in agreement with the standard-model values[2].

While our preferred result is based on the data-driven analysis of PVES fit, the final determination of the weak charge of the proton does not change appreciably with additional theoretical constraints. One of the dominant uncertainties in the term $B(Q^2, \theta)$ of equation (3) arises from the knowledge of the strange-quark contributions. These have been determined very precisely in recent theoretical calculations[16,17] employing lattice quantum chromodynamics (LQCD). Using these theoretical results to constrain the extrapolation to $Q^2 = 0$ results in a slightly lower weak charge and a reduction in the uncertainty, as shown in Table 1 ('PVES fit + LQCD'). The APV result was not included in this determination of Q_w^p; its inclusion makes negligible difference.

Because the proximity to threshold ($Q^2 \to 0$) and precision of our Q_{weak} result overwhelmingly dominate the fits described above, it is possible to go one step further and use the Q_{weak} datum by itself to determine Q_w^p. The fact that the strange and axial form factors contribute so little at the kinematics of the Q_{weak} experiment (0.1% and 2.5%, respectively) also helps motivate this consistency check. Using the same electromagnetic form factors[9] as in the fits above, the same lattice calculation[16] for the strange form factors, and following the extraction method of ref. [18] for the axial form factor, the Q_w^p result obtained by using just the Q_{weak} datum falls in-between the consistent results of the other determinations described above, which employ the entire PVES database (see Table 1, 'Q_{weak} datum only'). The uncertainty of the Q_w^p result in this case includes an additional uncertainty (4.6 p.p.b.) due to the calculated form factors, but is only 4% larger than the uncertainty of the global fit result, which uses the entire PVES database. The dominant correction, from the electromagnetic form factors (23.7%), is well known in the low-Q^2 regime of the Q_{weak} experiment.

The Q_w^p determinations described above can be used to test the prediction of the standard model for $\sin^2\theta_W$, the fundamental

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

electroweak parameter characterizing the mixing of the electromagnetic and weak interactions in the standard model (θ_W is the Weinberg angle). By neglecting small quantum corrections, the standard model derives[19,20] $\sin^2\theta_W$ in terms of the electroweak boson masses, M_W and M_Z, as $\sin^2\theta_W = 1-(M_W/M_Z)^2 \approx 1/4$, and so $Q_w^p = 1-4\sin^2\theta_W$ is nearly zero. This 'accidental' standard-model suppression of Q_w^p makes it an ideal observable in the search for new parity-violation interactions of natural size[21]. Using the latest input[2] to calculate quantum corrections that relate[1] $\sin^2\theta_W(Q=0)$ to Q_w^p, as described in Methods, we obtain $\sin^2\theta_W(Q=0)_{\overline{MS}} = 0.2383 \pm 0.0011$ in the modified-minimal-subtraction ($\overline{MS}$) scheme[2,19].

Figure 3 shows our $\sin^2\theta_W$ result (after subtracting 0.00012 to plot it at the Q_{weak} energy scale, 0.158 GeV), along with other $\sin^2\theta_W$ determinations[2,20,22–25]. Our value is consistent with the standard-model expectation and the purely leptonic E158 result[22] obtained in Møller (ee) scattering, which has different sensitivities to new physics from our semi-leptonic (ep) result. Although the measurements at the Z^0 pole are more precise than our result, there exist a variety of beyond-standard-model (BSM) scenarios that can have considerable influence on low-energy precision measurements but little effect on collider measurements at the Z^0-mass energy scale[26]. An example is the dark-photon model of ref. [27], which allows large effects for dark Z mediators of a few hundred megaelectronvolts at low Q, but no effects at the Z^0 pole.

To explore this experiment's sensitivity to new BSM contact interactions, we follow the convention[28] where a 'new physics' term, g^2/Λ^2, is added to the standard-model term $g_{AV}^{eq}/(2v^2)$ used in the Lagrangian to describe the neutral-current interaction of axial-vector electrons with vector quarks. (This convention differs from an earlier one[1] by a factor of 4; a factor of 2 in Λ). Here $g_{AV}^{eq} = C_{1q} = 2g_A^e g_V^q$ is the standard-model axial-electron–vector-quark coupling, $v^2 = \sqrt{2}/(2G_F)$, and Λ represents the mass reach for new physics (the mass of the hypothetical BSM particle being exchanged) with coupling g. Expressed in terms of Q_w^p and its uncertainty, $\pm\Delta Q_w^p$, the 95%-confidence-level mass reach is

$$\frac{\Lambda_\pm}{g} = v\sqrt{\frac{4\sqrt{5}}{|Q_w^p \pm 1.96\Delta Q_w^p - Q_w^p(\mathrm{SM})|}} \qquad (4)$$

where $\Lambda_+/g = 7.4$ TeV and $\Lambda_-/g = 8.4$ TeV.

For the extreme contact-interaction coupling[29] $g^2 = 4\pi$, the maximum mass reach for new semi-leptonic physics determined by our Q_w^p result is $\Lambda_+ = 26.3$ TeV. Using the coupling $g^2 = 4\pi\alpha$, which is typically assumed for leptoquarks[30] (hypothetical BSM particles with both lepton and baryon numbers; see also http://pdglive.lbl.gov/DataBlock.action?node=S056ESP&init=0), equation (4) rules out leptoquark masses below 2.3 TeV.

In a more general case[8] where the coupling of the new physics to the quarks is not restricted to the same 2:1 ratio of u to d quarks as in the proton, the mass reach can be expressed as a circle of radius $(g/\Lambda)^2\sqrt{2}/G_F$, about the standard-model origin $(C_{1u}^{SM}, C_{1d}^{SM}) = (-0.1887, 0.3419)$ in C_{1q} space. This is illustrated in Fig. 4a, which also shows the complementary constraints on C_{1q} provided by the weak-charge measurements of this experiment ($Q_w^p = -2(2C_{1u} + C_{1d})$) and the ^{133}Cs APV[14,15] result[2] ($Q_w^{Cs} = -2[55(2C_{1u} + C_{1d}) + 78(C_{1u} + 2C_{1d})]$). The isospin dependence of potential BSM physics can be expressed in terms of the projections $h_v^u = \cos\theta_h$ and $h_v^d = \sin\theta_h$, where θ_h is the quark-flavour mixing angle, the superscripts u and d refer to the quark flavour and the subscript refers to the vector piece, as in ref. [8]. We then obtain the 95%-confidence-level bounds from this experiment, which are shown in Fig. 4b, together with those from APV[2,14,15] and from combining those two constraints. New physics is ruled out below the curves. Our combined constraint raises the θ_h-independent limit for generic new semi-leptonic parity-violation BSM physics to 3.5 TeV.

The low-energy precision frontier continues to offer an exciting landscape to search for BSM physics. The results of this experiment are consistent with the standard model and place important limits on new beyond-standard-model physics. Future experiments propose to provide even more precise (and much more challenging) determinations of Q_w^p at lower Q^2 (ref. [31]) and of Q_w^e (ref. [32]) by taking the techniques and lessons learned in this experiment to the next level.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0096-0.

Received: 12 October 2017; Accepted: 5 April 2018;
Published online 9 May 2018.

1. Erler, J., Kurylov, A. & Ramsey-Musolf, M. J. Weak charge of the proton and new physics. *Phys. Rev. D* **68**, 016006 (2003).
2. Particle Data Group. Review of particle physics. *Chinese Phys. C* **40**, 100001 (2016).
3. Androić, D. et al. First determination of the weak charge of the proton. *Phys. Rev. Lett.* **111**, 141803 (2013).
4. Allison, T. et al. The Q_{weak} experimental apparatus. *Nucl. Instrum. Meth. A* **781**, 105–133 (2015).
5. Narayan, A. et al. Precision electron-beam polarimetry at 1 GeV using diamond microstrip detectors. *Phys. Rev. X* **6**, 011013 (2016).
6. Smith, G. R. The Q_{weak} target performance. *Nuovo Cim. C* **35**, 159–163 (2012).
7. Young, R. D., Roche, J., Carlini, R. D. & Thomas, A. W. Extracting nucleon strange and anapole form factors from world data. *Phys. Rev. Lett.* **97**, 102002 (2006).
8. Young, R. D., Carlini, R. D., Thomas, A. W. & Roche, J. Testing the standard model by precision measurement of the weak charges of quarks. *Phys. Rev. Lett.* **99**, 122003 (2007).
9. Arrington, J. & Sick, I. Precise determination of low-Q nucleon electromagnetic form factors and their impact on parity-violating e-p elastic scattering. *Phys. Rev. C* **76**, 035201 (2007).
10. Hall, N. L., Blunden, P. G., Melnitchouk, W., Thomas, A. W. & Young, R. D. Quark-hadron duality constraints on γZ box corrections to parity-violating elastic scattering. *Phys. Lett. B* **753**, 221–226 (2016).
11. Blunden, P. G., Melnitchouk, W. & Thomas, A. W. New formulation of γZ box corrections to the weak charge of the proton. *Phys. Rev. Lett.* **107**, 081801 (2011).
12. Blunden, P. G., Melnitchouk, W. & Thomas, A. W. γZ box corrections to weak charges of heavy nuclei in atomic parity violation. *Phys. Rev. Lett.* **109**, 262301 (2012).
13. Gorchtein, M., Horowitz, C. J. & Ramsey-Musolf, M. J. Model dependence of the γZ dispersion correction to the parity-violating asymmetry in elastic ep scattering. *Phys. Rev. C* **84**, 015502 (2011).
14. Wood, C. S. et al. Measurement of parity nonconservation and an anapole moment in cesium. *Science* **275**, 1759–1763 (1997).
15. Dzuba, V. A., Berengut, J. C., Flambaum, V. V. & Roberts, B. Revisiting parity nonconservation in cesium. *Phys. Rev. Lett.* **109**, 203003 (2012).
16. Green, J. et al. High-precision calculation of the strange nucleon electromagnetic form factors. *Phys. Rev. D* **92**, 031501 (2015).
17. Sufian, R. S. et al. Strange quark magnetic moment of the nucleon at the physical point. *Phys. Rev. Lett.* **118**, 042001 (2017).
18. Liu, J., McKeown, R. D. & Ramsey-Musolf, M. J. Global analysis of nucleon strange form factors at low Q^2. *Phys. Rev. C* **76**, 025202 (2007).
19. Erler, J. & Ramsey-Musolf, M. J. Weak mixing angle at low energies. *Phys. Rev. D* **72**, 073003 (2005).
20. Kumar, K. S., Mantry, S., Marciano, W. J. & Souder, P. A. Low-energy measurements of the weak mixing angle. *Annu. Rev. Nucl. Part. Sci.* **63**, 237–267 (2013).
21. Seiberg, N. Naturalness versus supersymmetric non-renormalization theorems. *Phys. Lett. B* **318**, 469–475 (1993).
22. Anthony, P. L. et al. Precision measurement of the weak mixing angle in Møller scattering. *Phys. Rev. Lett.* **95**, 081601 (2005).
23. The Jefferson Lab PVDIS Collaboration. Measurement of parity violation in electron-quark scattering. *Nature* **506**, 67–70 (2014).
24. The NuTeV Collaboration. Precise determination of electroweak parameters in neutrino–nucleon scattering. *Phys. Rev. Lett.* **88**, 091802 (2002); errratum 90, 239902 (2003).
25. Bentz, W., Cloët, I. C., Londergan, J. T. & Thomas, A. W. Reassessment of the NuTeV determination of the weak mixing angle. *Phys. Lett. B* **693**, 462–466 (2010).
26. Erler, J. & Su, S. The weak neutral current. *Prog. Part. Nucl. Phys.* **71**, 119–149 (2013).
27. Davoudiasl, H., Lee, H.-S. & Marciano, W. J. Muon $g-2$, rare kaon decays, and parity violation from dark bosons. *Phys. Rev. D* **89**, 095006 (2014).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

28. Erler, J., Horowitz, C. J., Mantry, S. & Souder, P. A. Weak polarized electron scattering. *Annu. Rev. Nucl. Part. Sci.* **64**, 269–298 (2014).
29. Eichten, E., Lane, K. D. & Peskin, M. E. New tests for quark and lepton substructure. *Phys. Rev. Lett.* **50**, 811–814 (1983).
30. Abramowicz, H. et al. Search for first-generation leptoquarks at HERA. *Phys. Rev. D* **86**, 012005 (2012).
31. Berger, N. et al. Measuring the weak mixing angle with the P2 experiment at MESA. *J. China Univ. Sci. Tech.* **46**, 481–487 (2016).
32. Benesch, J. et al. The MOLLER experiment: an ultra-precise measurement of the weak mixing angle using Møller scattering. Preprint at https://arxiv.org/abs/1411.4088v2 (2014).

Acknowledgements This work was supported by the US Department of Energy (DOE) Contract number DEAC05-06OR23177, under which Jefferson Science Associates, LLC, operates the Thomas Jefferson National Accelerator Facility. Construction and operating funding for the experiment was provided through the US DOE, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Foundation for Innovation (CFI), and the National Science Foundation (NSF) with university matching contributions from the College of William and Mary, Virginia Tech, George Washington University and Louisiana Tech University. We thank the staff of Jefferson Laboratory, in particular the accelerator operations staff, the target and cryogenic groups, the radiation control staff, as well as the Hall C technical staff for their help and support. We are grateful for the contributions of our undergraduate students. We thank TRIUMF for its contributions to the development of the spectrometer and integrated electronics, and BATES for its contributions to the spectrometer and Compton polarimeter. We are indebted to P. G. Blunden, J. D. Bowman, J. Erler, N. L. Hall, W. Melnitchouk, M. J. Ramsey-Musolf and A. W. Thomas for discussions. We also thank P. A. Souder for contributions to the analysis. Figure 2 was adapted with permission from ref. [3] (copyrighted by the American Physical Society).

Author contributions Authors contributed to one or more of the following areas: proposing, leading and running the experiment; design, construction, optimization and testing of the experimental apparatus and data acquisition system; data analysis; simulation; extraction of the physics results from measured asymmetries; and the writing of this Letter.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0096-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to G.R.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The Jefferson Lab Q_{weak} Collaboration

D. Androić[1], D. S. Armstrong[2], A. Asaturyan[3], T. Averett[2], J. Balewski[4], K. Bartlett[2], J. Beaufait[5], R. S. Beminiwattha[6], J. Benesch[5], F. Benmokhtar[7], J. Birchall[8], R. D. Carlini[2,5]*, J. C. Cornejo[2], S. Covrig Dusa[5], M. M. Dalton[9], C. A. Davis[10], W. Deconinck[2], J. Diefenbach[11], J. F. Dowd[2], J. A. Dunne[12], D. Dutta[12], W. S. Duvall[13], M. Elaasar[14], W. R. Falk[8,25], J. M. Finn[2,26], T. Forest[15,16], C. Gal[9], D. Gaskell[5], M. T. W. Gericke[8], J. Grames[17], V. M. Gray[2], K. Grimm[2,16], F. Guo[4], J. R. Hoskins[2], D. Jones[9], M. Jones[5], R. Jones[18], M. Kargiantoulakis[9], P. M. King[6], E. Korkmaz[19], S. Kowalski[4], J. Leacock[13], J. Leckey[2], A. R. Lee[13], J. H. Lee[2,6], L. Lee[8,10], S. MacEwan[8], D. Mack[5], J. A. Magee[2], R. Mahurin[8], J. Mammei[13], J. W. Martin[20], M. J. McHugh[21], D. Meekins[5], J. Mei[5], K. E. Mesick[21], R. Michaels[5], A. Micherdzinska[21], A. Mkrtchyan[3], H. Mkrtchyan[3], N. Morgan[13], A. Narayan[12], L. Z. Ndukum[12], V. Nelyubin[9], H. Nuhait[16], Nuruzzaman[11,12], W. T. H. van Oers[8,10], A. K. Opper[21], S. A. Page[8], J. Pan[8], K. D. Paschke[9], S. K. Phillips[22], M. L. Pitt[13], M. Poelker[17], J. F. Rajotte[4], W. D. Ramsay[8,10], J. Roche[6], B. Sawatzky[5], T. Seva[1], M. H. Shabestari[12], R. Silwal[9], N. Simicevic[16], G. R. Smith[5]*, P. Solvignon[5,26], D. T. Spayde[23], A. Subedi[12], R. Subedi[21], R. Suleiman[17], V. Tadevosyan[3], W. A. Tobias[9], V. Tvaskis[8,20], B. Waidyawansa[6], P. Wang[8], S. P. Wells[16], S. A. Wood[5], S. Yang[2], R. D. Young[24], P. Zang[25] & S. Zhamkochyan[3]

[1]Department of Physics, University of Zagreb, Zagreb, Croatia. [2]Department of Physics, College of William and Mary, Williamsburg, VA, USA. [3]Division of Experimental Physics, A. I. Alikhanyan National Science Laboratory (Yerevan Physics Institute), Yerevan, Armenia. [4]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Physics Division, Thomas Jefferson National Accelerator Facility, Newport News, VA, USA. [6]Department of Physics and Astronomy, Ohio University, Athens, OH, USA. [7]Department of Physics, Christopher Newport University, Newport News, VA, USA. [8]Department of Physics and Astronomy, University of Manitoba, Winnipeg, Manitoba, Canada. [9]Department of Physics, University of Virginia, Charlottesville, VA, USA. [10]Science Division, TRIUMF, Vancouver, British Columbia, Canada. [11]Department of Physics, Hampton University, Hampton, VA, USA. [12]Department of Physics and Astronomy, Mississippi State University, Mississippi State, MS, USA. [13]Department of Physics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. [14]Department of Natural Sciences, Southern University at New Orleans, New Orleans, LA, USA. [15]Department of Physics, Idaho State University, Pocatello, ID, USA. [16]Department of Physics, Louisiana Tech University, Ruston, LA, USA. [17]Accelerator Division, Thomas Jefferson National Accelerator Facility, Newport News, VA, USA. [18]Department of Physics, University of Connecticut, Storrs-Mansfield, CT, USA. [19]Department of Physics, University of Northern British Columbia, Prince George, British Columbia, Canada. [20]Department of Physics, University of Winnipeg, Winnipeg, Manitoba, Canada. [21]Department of Physics, George Washington University, Washington DC, USA. [22]Department of Physics, University of New Hampshire, Durham, NH, USA. [23]Physics Department, Hendrix College, Conway, AR, USA. [24]Department of Physics and Mathematical Physics, University of Adelaide, Adelaide, South Australia, Australia. [25]Department of Physics, Syracuse University, Syracuse, NY, USA. [26]Deceased: W. R. Falk, J. M. Finn, P. Solvignon.
*e-mail: carlini@jlab.org; smithg@jlab.org

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Here we describe the formalism connecting the experimentally measured asymmetry A_{ep} to the proton's weak charge Q_w^p, the experimental method used to determine A_{ep}, some measures of the data quality, and the electroweak radiative corrections needed to extract the weak mixing angle from Q_w^p.

Formalism. The Q_{weak} experiment measured the parity-violating asymmetry A_{ep}, which is the normalized difference between the elastic-scattering cross-sections (σ) of longitudinally polarized electrons with positive (σ_+) and negative (σ_-) helicity from unpolarized protons:

$$A_{ep} = \frac{\sigma_+ - \sigma_-}{\sigma_+ + \sigma_-} \tag{5}$$

The extended structure of the proton can be represented using form factors. Assuming charge symmetry, the single-boson exchange ('tree-level') expression for the asymmetry can be written in terms of the proton and neutron electromagnetic form factors G_E^p, G_M^p, G_E^n and G_M^n, the strange electric and magnetic form factors G_E^s and G_M^s, and the neutral weak axial form factor G_A^p as[7,33]

$$A_{ep} = A_0 \frac{A_v^p + A_s^p + A_a^p}{\varepsilon (G_E^p)^2 + \tau (G_M^p)^2} \tag{6}$$

with

$$\begin{aligned} A_0 &= \frac{-G_F Q^2}{4\pi\alpha\sqrt{2}} \\ A_v^p &= Q_w^p[\varepsilon (G_E^p)^2 + \tau (G_M^p)^2] \\ &\quad - [\varepsilon G_E^p G_E^n + \tau G_M^p G_M^n] \\ A_s^p &= -\varepsilon G_E^p G_E^s - \tau G_M^p G_M^s \\ A_a^p &= -(1 - 4\sin^2\theta_W)\varepsilon' G_M^p G_A^p \\ Q_w^p &= -2(2C_{1u} + C_{1d}) \end{aligned}$$

where the subscripts v, s and a refer to the vector, strange and axial contributions to the asymmetry, respectively,

$$\varepsilon = \frac{1}{1 + 2(1+\tau)\tan^2(\theta/2)} \quad \text{and} \quad \varepsilon' = \sqrt{\tau(1+\tau)(1-\varepsilon^2)}$$

are kinematic quantities and $\tau = Q^2/(4M^2)$, where M is the proton mass.

By taking the limits $\theta \to 0$ and $Q^2 \to 0$, equation (6) is reduced[34] to equation (3). Although similar relationships apply for parity-violating elastic scattering from ^{4}He and quasi-elastic scattering from ^{2}H, these involve different linear combinations of C_{1u} and C_{1d} and the various form factors[33,35].

Global fit. Using equation (6), a global fit was made to a database that included all relevant PVES data up to $Q^2 = 0.63\ \mathrm{GeV}^2\,c^{-2}$: 28 electron–proton scattering results obtained by the G0[36,37], HAPPEX[38–41], SAMPLE[42], PVA4[43–45] and Q_{weak}[3] collaborations, including the present result, two ^{4}He elastic scattering results (HAPPEX[40,46]) and five ^{2}H quasi-elastic scattering results (G0[37], PVA4[45,47] and SAMPLE[48]).

For the fit, we followed the procedure introduced by Young et al.[7], as used in refs [3,8]. We used six parameters in the fit; namely, C_{1u}, C_{1d} and four parameters that characterize the strange and axial form factors: the strange radius ρ_s, the strange magnetic moment μ_s and the magnitudes of G_A^p and G_A^n. The parameterizations chosen for the strange form factors were $G_E^s = \rho_s Q^2 G_d$ and $G_M^s = \mu_s G_d$, with the dipole form factor $G_d = (1 - Q^2/\lambda^2)^{-2}$, where $\lambda = 1\ \mathrm{GeV}\,c^{-1}$. Young et al.[7] examined the consequences of using more elaborate Q^2 dependences for G_E^s and G_M^s and found that they were not required for the data. More recent analyses have come to the same conclusion[18,49]. LQCD calculations[16,17] have also found similar shapes for G_E^s and G_M^s. The neutron's axial form factor G_A^n enters the fit through the inclusion of ^{2}H data in the database. For the isovector combination $G_A^{Z(T=1)} = (G_A^p - G_A^n)/2$, the dipole form G_d was again adopted, with the normalization factor as the parameter. Because the isoscalar combination $G_A^{Z(T=0)} = (G_A^p + G_A^n)/2$ is known from the theory to be small, it was constrained in the fit to the theoretical value[50] $G_A^{Z(T=0)} = -0.08 \pm 0.26$, reducing the effective number of parameters to five. The unconstrained isovector combination $G_A^{Z(T=1)}$ presented in Table 1 was constructed from the values of G_A^p and G_A^n determined in the fit, which were $G_A^p = -0.59 \pm 0.34$ and $G_A^n = 0.68 \pm 0.44$ with a covariance of -0.0265.

The electromagnetic form factors G_E^p, G_M^p, G_E^n and G_M^n were taken from ref. [9]. If, instead, we use any one of the several alternative parameterizations for these form factors[51–53], the fitted result for Q_w^p changes by less than 1%. We have incorporated this range as a systematic uncertainty.

Each of the experimental asymmetries A_{ep} used in the fit needed to be corrected for the one electroweak radiative correction with considerable energy dependence, the γZ-box diagram[1]. Three independent theoretical groups have calculated this correction, with results in excellent mutual agreement[10,13,54]. We have adopted the most recent, which was obtained from a data-driven calculation of the vector[10] (0.0054(4)) and axial-vector[11,12] (−0.0007(2)) contributions, multiplied by a small Q^2 correction[13] (0.978(12)); the numerical values given here are in terms of the effect on the reduced asymmetry A_{ep}/A_0. The total correction to our datum is 0.0046(5), corresponding to a 6.4% ± 0.6% correction to Q_w^p. If, instead, we use the calculations of either ref. [13] or ref. [54], the extracted value of Q_w^p is essentially unchanged, and the effect on its uncertainty is negligible.

The fit was done using linear χ^2 minimization, with a resulting $\chi^2/\mathrm{d.o.f.} = 1.25$ for 29 degrees of freedom (d.o.f.).

Apparatus. The Q_{weak} experiment was performed using a custom apparatus installed in Hall C at the Thomas Jefferson National Accelerator Facility. A complete description of the apparatus and critical aspects of the accelerator can be found in ref. [4]. Here we highlight some of the most important details illustrated in Extended Data Fig. 1.

The electron beam was longitudinally polarized and its helicity was reversed at a rate of 960 Hz in a pseudorandom sequence of 'helicity quartets' (+ − − +) or (− + + −). The quartet pattern minimized noise due to slow linear drifts, and the rapid helicity reversal limited noise due to fluctuations in the target density and beam properties. As a test for possible false asymmetries, two methods were used to reverse the beam helicity on a slower timescale than the rapid helicity reversal. About every 8 h, the helicity of the laser beam that was used to generate the polarized electron beam was reversed by insertion of a half-wave plate in its path. The helicity of the electron beam in the accelerator's injector region was reversed monthly using a 'double Wien' spin rotator[55]. Non-vanishing correlations between the properties of the electron beam (intensity, position, angle and energy) and its helicity lead to false contributions to the measured asymmetry. These helicity-correlated beam properties were suppressed by carefully setting up the helicity-defining optics and active-feedback systems in the polarized injector laser system. The electron beam transport line was instrumented to allow the correction of residual helicity-correlated beam properties. Beam monitors upstream of the experimental apparatus provided continuous, non-invasive measurement of the electron beam's intensity, position, angle and energy. The response of the experimental apparatus to changes in the beam properties was periodically measured using a beam modulation system that generated controlled variations in the beam's position and angle using magnets and in its energy using a radio-frequency accelerating cavity. Finally, the electron beam's polarization was measured with two independent polarimeters upstream of the Q_{weak} apparatus.

The experiment usually ran with a 180-μA beam of 1.16-GeV, ∼88% longitudinally polarized electrons incident on a 34.4-cm-long liquid hydrogen target placed in an aluminium cell and maintained at 20 K. A set of three lead collimators restricted the polar scattering angle acceptance to the range $5.8° \le \theta \le 11.6°$ with an azimuthal angle coverage of 49% of 2π. A resistive toroidal magnet between the target and detectors separated the elastically scattered electrons of interest from irreducible backgrounds arising from inelastic and Møller scattering. The resulting scattered electrons were detected in eight synthetic-quartz Cherenkov detectors, each 200 cm × 18 cm × 1.25 cm, arranged in an azimuthally symmetric pattern about the beam axis. Fourfold azimuthal symmetry was important to minimize and characterize the effects of helicity-correlated beam properties and residual transverse beam polarization. These 'main' detectors were equipped with 2-cm-thick Pb pre-radiators that amplified the electron signal and suppressed low-energy backgrounds. Light was collected from the detectors by photomultiplier tubes (PMTs) at each end. The high rates of ∼800 MHz per detector required a current-mode readout, in which the PMT anode current was converted to a voltage that was digitized and integrated about every millisecond. Photographs of the apparatus are shown in Extended Data Fig. 1.

The collimator–magnet configuration and carefully designed shielding were effective in minimizing reducible sources of background, such as direct line-of-sight (neutral) tracks originating in the target and secondary scattering from the beampipe. The residual diffuse background was monitored by background detectors in the main-detector shield house. These consisted of a complete main detector placed in the super-elastic region, a dark box with a bare PMT and a dark box with a PMT attached to a light guide. A symmetric array of four smaller detectors ('upstream luminosity monitors') placed on the upstream face of the defining (middle) collimator was effective in monitoring residual backgrounds from the tungsten beam collimator that shielded the downstream region from small-angle-scattered particles.

Finally, a tracking system consisting of drift chambers located before and after the magnet was deployed periodically to verify the acceptance-weighted kinematic distribution and to help study the backgrounds. These measurements were done in special run periods at very low beam currents (0.1–200 nA) and using conventional individual pulse counting.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Data from the experiment's short commissioning run (4% of the size of the dataset reported here) have been previously published[3]. Here we report the combined result from two run periods (referred to as Run 1 and Run 2), each about six months in duration. Because improvements were made to the apparatus between the two run periods and their beam conditions were different, we report the corrections and systematic uncertainties for each run period separately. To prevent possible biases in the analysis, the main-detector asymmetries were blinded by an additive shift in the asymmetry that was different for each run period. When the data analysis was complete, the asymmetries were unblinded, revealing the results presented here.

Data analysis. The raw asymmetry A_{raw} was formed from the normalized cross-section difference of equation (5) over the sum of the beam-charge normalized detector PMT signals, summed over the 8 detectors. The measured asymmetry A_{msr} was calculated from A_{raw} by correcting for a variety of effects that could cause false asymmetries:

$$A_{\mathrm{msr}} = A_{\mathrm{raw}} + A_{\mathrm{t}} + A_{\mathrm{l}} + A_{\mathrm{bcm}} + A_{\mathrm{bb}} + A_{\mathrm{beam}} + A_{\mathrm{bias}} \quad (7)$$

The methods for determining each of these corrections and their uncertainties are discussed below.

Transverse asymmetry correction, A_{t}. A small (∼2%) residual transverse component of the incident beam polarization resulted in a transverse asymmetry A_{t}, which is driven by a parity-conserving two-photon-exchange amplitude[56]. To determine this correction, we measured the transverse asymmetry using a maximally transversely polarized incident beam[57], which we used to calculate an upper bound on the broken symmetry of the spectrometer–detector system (<1.3%). By combining these effects with the measured residual transverse polarization components of the beam during data-taking with a nominally longitudinally polarized beam, we determined the correction to be $A_{\mathrm{t}} = 0.0 \pm 1.1$ p.p.b. and $A_{\mathrm{t}} = 0.0 \pm 0.7$ p.p.b. for Run 1 and Run 2, respectively.

Linearity correction, A_{l}. The nonlinearity of the Cherenkov detector readout chain (PMT, low-noise voltage-to-current preamplifier and analogue-to-digital converter) was directly measured using a system with light-emitting diodes and was found to be $0.7\% \pm 0.5\%$ at signal levels corresponding to those recorded during the experiment. This results in an additive correction $A_{\mathrm{l}} = 1.3 \pm 1.0$ p.p.b. and $A_{\mathrm{l}} = 1.2 \pm 0.9$ p.p.b. for Run 1 and Run 2, respectively.

Beam current monitor correction, A_{bcm}. The beam current was measured non-invasively with radio-frequency resonant cavities to allow a precise relative comparison of the beam charge in each helicity state. Two such beam current monitors (BCMs) were used in Run 1, while three BCMs were used in Run 2 after the installation of an additional monitor and improvement of the low-noise digital demodulation electronics. The correction A_{bcm} is zero by definition because in each period we normalize the integrated detector signals to the average of the BCM signals. The systematic uncertainty on this correction is determined from the variation in the reported charge asymmetry for the BCMs in each analysis, resulting in $\delta A_{\mathrm{bcm}} = \pm 4.4$ p.p.b. and $\delta A_{\mathrm{bcm}} = \pm 2.1$ p.p.b. for Run 1 and Run 2, respectively.

Beamline background asymmetry correction, A_{bb}. False asymmetry caused by secondary events scattered from the beamline and the tungsten beam collimator is referred to as the beamline background asymmetry A_{bb}. Such events were determined to be predominantly due to low-energy neutral particles and contributed a small amount to the signal (0.19%) but had a large asymmetry, which is believed to be associated with a helicity-dependent intensity or position variation in the extended halo around the main accelerated beam. Although their contribution was only a small component of the main detector signal, it dominated the asymmetry measurement of the background detectors, which were highly correlated (see Extended Data Fig. 2b). A direct correlation between the main-detector asymmetry from these events and the background asymmetries measured by the background detectors was shown by blocking two of the eight openings in the first of the three Pb collimators with 5.1-cm-thick tungsten plates (see Extended Data Fig. 2a). To correct for this false asymmetry, a correlation factor was extracted (separately for Run 1 and Run 2) between the asymmetries of the main detector array and the upstream luminosity monitor array, as shown in Extended Data Fig. 2c. The correlation factor was combined with the measured upstream luminosity monitor asymmetry, averaged every 6 min, to correct the main-detector asymmetry in that interval. The resulting net corrections for Run 1 and Run 2 were $A_{\mathrm{bb}} = 3.9 \pm 4.5$ p.p.b. and $A_{\mathrm{bb}} = -2.4 \pm 1.1$ p.p.b., respectively, where the uncertainty includes contributions from the statistical error on the determination of the correlation, as well as systematic errors extracted by allowing the correlation factor to vary randomly within a reasonable range over different time periods[58].

Correction for helicity-correlated beam properties, A_{beam}. Residual non-vanishing correlations in the properties of the electron beam were accounted for through the correction A_{beam}. Five beam properties—the transverse beam position and angle (horizontal X, X'; vertical Y,Y'), as well as the energy—were monitored continuously as described above. The run-averaged helicity-correlated values of these parameters are presented in Extended Data Table 1. Using the measured helicity-correlated beam differences $\Delta\chi_i$ and the sensitivity of the measured asymmetry to variations in the beam parameters $\partial A/\partial\chi_i$, the corrections for the ith beam parameter were combined into the total correction:

$$A_{\mathrm{beam}} = -\sum_{i=1}^{5} \left(\frac{\partial A}{\partial \chi_i} \right) \Delta\chi_i$$

The detector sensitivities $\partial A/\partial\chi_i$ were measured routinely by varying the beam parameters using the beam modulation system described above. This system created small perturbations of the beam parameters about their nominal values with a sinusoidal driving function at 125 Hz. Furthermore, an accelerator fast-feedback system with an operational frequency range that included 125 Hz was used, driving the beam position and angle parameters with a sinusoidal pattern that was 90° out of phase with respect to the beam modulation system. The result was that the beam parameters were driven in more than one way, allowing for redundant measurements of the detector sensitivities using different combinations of the driven signals. The spread in those results was the dominant component of the systematic uncertainty for this correction. Typical values of the sensitivities are shown in Extended Data Table 1. For a perfectly symmetric apparatus, the position and angle sensitivities would be zero. For this apparatus, the sensitivities show that the horizontal plane had a larger broken symmetry than the vertical plane. The resulting corrections were $A_{\mathrm{beam}} = 18.5 \pm 4.1$ p.p.b. and $A_{\mathrm{beam}} = 0.0 \pm 1.1$ p.p.b. for Runs 1 and 2, respectively. The considerably smaller correction in Run 2 was due to smaller position and angle helicity-correlated differences during that period.

Rescattering bias, A_{bias}. When comparing the measured asymmetry in the two PMTs ('left' and 'right') that read out the signal at each end of the eight main detectors, we found a consistent difference of about +300 p.p.b., as shown in Extended Fig. 3a. Here, 'right' is the beam direction $\hat{\boldsymbol{k}}$ crossed with the radial direction $\hat{\boldsymbol{r}}$, $\hat{\boldsymbol{k}} \times \hat{\boldsymbol{r}}$. This effect is due to the left/right analysing power of the multiple scattering of transversely (radially) polarized electrons through the Pb pre-radiators of the main detectors. For perfect symmetry, this parity-conserving effect cancels when forming the parity-violating asymmetry of interest. Properly accounting for the minor broken symmetries of the as-built apparatus leads to a correction, A_{bias}, referred to as rescattering bias.

A schematic of the physical model for this effect is shown in Extended Data Fig. 3a. Scattered electrons, which are initially fully longitudinally polarized, acquire some transverse polarization through precession as they transport through the spectrometer's magnetic field.

The effect was modelled using a detailed Geant4[59] simulation of the transport of the detected electrons through the spectrometer and the Pb radiator, including electromagnetic showering and Mott scattering. An asymmetry in the distribution of electrons penetrating the radiator develops owing to the analysing power of low-energy Mott scattering. A possible analysing power for high-energy scattering due to non-Born processes[60] was also considered, but reasonable models for those processes[61] showed a negligible contribution.

Asymmetries obtained from this simulation (and from a variety of analytic effective models that reproduced the key features of the simulation) were combined with scattered electron flux distributions and tailored parameterizations of the Cherenkov-light yield for each detector to estimate A_{bias} and its uncertainty. The light yield varied strongly with the arrival angle and position of the electron on the detector. The light-yield parameterizations were developed to match the observed light-yield distribution by tuning the optical parameters in a Geant4[59] optical-photon transport simulation. The largest systematic uncertainty was associated with the optical modelling of the individual as-built detectors. This uncertainty was determined from the range of predicted A_{bias} values, which was obtained by varying the optical parameters in the simulation while maintaining reasonable agreement with the measured light yield distributions. The predicted A_{bias} values for each detector are shown in Extended Data Fig. 3c. The resulting averaged correction and its systematic uncertainty are $A_{\mathrm{bias}} = 4.3 \pm 3.0$ p.p.b.

This result was consistent to within 1 p.p.b. with a simpler, independent calculation based on a phenomenological approach, which used the measured position distributions of the electron flux on the pre-radiators and the measured dependence of the light seen by each PMT on the track position on the detectors. The results obtained using the measured distributions from the experiment's tracking system were compared with those obtained from the simulation. In this approach, the effect of the position differences was scaled to match the observed left/right asymmetry. The sensitivity to various models for the position dependence of the asymmetries was found to be small.

Determination of A_{ep}. The measured asymmetry A_{msr} was then corrected for incomplete beam polarization, the effects of various background processes, electromagnetic radiative corrections and the finite acceptance of the detector, to obtain the fully corrected electron–proton asymmetry A_{ep}, using:

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

$$A_{ep} = R_{tot}\frac{A_{msr}/P - \sum_{i=1,3,4} f_i A_i}{1 - \sum_{i=1}^{4} f_i} \quad (8)$$

where $R_{tot} = R_{rc}R_{det}R_{acc}R_{Q^2}$, f_i are dilutions to the signal and A_i are false or background process asymmetries. The components of equation (8) are discussed below.

(1) R_{rc}. The electromagnetic radiative correction factor $R_{rc} = 1.010 \pm 0.005$ accounts for the effect of internal and external bremsstrahlung of the incident electron, which can depolarize the electron and modify the momentum transfer Q^2 at the scattering vertex. R_{rc} was determined using a Geant3[59] simulation by comparing results with and without bremsstrahlung enabled in the simulation.

(2) R_{det}. The Cherenkov detector analogue response (that is, the summed optical signal detected by the two PMTs attached at each end of each detector) varied as a function of the arrival location of the scattered electron on the detector. The magnetic optics of the spectrometer also caused a correlation of the electron arrival location with Q^2 and, therefore, with the asymmetry. The correlation between the detector's analogue response and the Q^2 value of each track was determined using the tracking-system drift chambers, and the resulting correction to the measured asymmetry was $R_{det} = 0.9895 \pm 0.0021$.

(3) R_{acc}. Owing to the finite acceptance of the spectrometer and the effect of radiative energy losses, A_{msr} represents an average over a range of Q^2 values. Because the asymmetry varies strongly with Q^2, we used a simulation to correct the averaged asymmetry $\langle A(Q^2)\rangle$ to the asymmetry that would arise from point scattering at the central $\langle Q^2\rangle$, $A(\langle Q^2\rangle)$, using:

$$R_{acc} = \frac{A(\langle Q^2\rangle)}{\langle A(Q^2)\rangle} = 0.977 \pm 0.002$$

(4) R_{Q^2}. The central $\langle Q^2\rangle$ for the experiment was determined from a Geant4 simulation that was benchmarked with measurements from the tracking system. The $\langle Q^2\rangle$ value was not identical for Run 1 and Run 2 because of minor differences in the beam energy, target location and magnetic field of the spectrometer, with Run 1 having a higher $\langle Q^2\rangle$. The global fit of A_{ep} versus Q^2 (see Fig. 2) was used to determine the sensitivity of the asymmetry to small changes in $\langle Q^2\rangle$. Run 2 was chosen as the reference for $\langle Q^2\rangle$, and the Run 1 asymmetry was scaled from its measured $\langle Q^2\rangle$ value using $R_{Q^2} = 0.9928$ ($R_{Q^2} = 1$ for Run 2, by definition). The determination of the central $\langle Q^2\rangle$ value has a 0.45% relative uncertainty, dominated by the uncertainties on the locations of the collimator, target and main detector and that of the beam energy determination. To simplify the global fitting, we decided to quote $\langle Q^2\rangle$ as exact and used the sensitivity $\partial A_{ep}/\partial Q^2$ to determine an effective error contribution to the asymmetry. This error on R_{Q^2} was 0.0055 for both run periods. The acceptance-averaged Q^2, scattering angle and incident electron energy were $\langle Q^2\rangle = 0.0248\ \mathrm{GeV}^2\,c^{-2}$, $\langle\theta\rangle = 7.90^\circ$ and $\langle E_0\rangle = 1.149$ GeV, respectively.

(5) Beam polarization, P. To achieve a reliable determination of the beam polarization (P) at <1% accuracy, two different techniques with precisely calculated analysing powers were employed for redundancy. An existing Møller polarimeter[62] in experimental Hall C was used invasively 2–3 times a week. It measured the parity-conserving cross-section asymmetry in the scattering of polarized beam electrons from polarized electrons in an iron foil target at low (typically smaller than or equal to ~2 μA) beam currents. A newly installed, non-invasive Compton polarimeter[7] monitored the beam polarization continuously at the full-production beam current of 180 μA. This device measured the parity-conserving asymmetry in the scattering of beam electrons from circularly polarized laser photons. For each run period, the averaged beam-polarization-corrected asymmetry was computed in two ways: by correcting each ~6-min period of data-taking for the polarization measured during that interval, and by using an overall average beam polarization for the whole run period. Because the two methods gave the same result to a small fraction of the quoted uncertainty, for simplicity the results obtained using the overall average beam polarization are quoted here. The overall average beam polarizations for the two running periods were $P_{Run1} = (87.66 \pm 1.05)\%$ and $P_{Run2} = (88.71 \pm 0.55)\%$, where the uncertainties are predominantly systematic. For Run 1, the uncertainty was larger for two reasons: the Compton polarimeter was still being commissioned, so it was not used for this determination, and the Møller uncertainty was larger than usual owing to the need to correct for the effects of an intermittent short circuit in one of the quadrupole magnets of the polarimeter. For Run 2, both polarimeters were fully functional and agreed well with each other, as shown in Extended Data Fig. 4. A dedicated direct comparison of the Møller and Compton polarimeters under identical beam conditions at low beam current was also performed. The two techniques agreed within the uncertainties for that measurement, $dP/P = 1\%$ and $dP/P = 0.73\%$ for the Compton and Møller polarimeters, respectively[63].

Physics backgrounds (target windows). The entrance (0.11-mm-thick) and exit (0.13-mm-thick) windows of the hydrogen target were made of aluminium 7075 alloy. Electrons scattered from these windows caused the dominant background process ($f_1A_1 = 37$ p.p.b. for Run 1 and $f_1A_1 = 38$ p.p.b. for Run 2). The parity-violating elastic asymmetry for electrons scattered from aluminium is observed to be nearly an order of magnitude larger than that for scattering from hydrogen owing to the much larger weak charge of the aluminium nucleus[64]. Therefore, even the small fraction of the detected yield arising from the windows required a substantial correction to the measured asymmetry. The aluminium asymmetry was determined in dedicated data-taking runs using an aluminium target made from the same block of material as the target windows but with a thickness (3.7 mm) to match the radiation length of the hydrogen target. The ranges of scattering angles accepted from the upstream and downstream windows were different, which required a small kinematic correction to the measured alloy asymmetry to yield the asymmetry from the target windows $A_1 = 1.515 \pm 0.077$ p.p.m. (see Extended Data Fig. 5). The uncertainty is dominated by statistics but includes systematic uncertainties arising from the kinematic correction, among others. The fraction of the measured yield arising from the target windows, $f_1 = (2.471 \pm 0.056)\%$ (Run 1) and $f_1 = (2.516 \pm 0.059)\%$ (Run 2), was measured with a low beam current on an evacuated target cell, and a simulation was used to correct for radiative effects due to the liquid hydrogen.

Physics backgrounds (beamline background dilution). As described above (see 'Beamline-background asymmetry correction, A_{bb}'), a component of the background came from scattering sources in the beamline. The dilution from this source was measured[58] to be $f_2 = (0.193 \pm 0.064)\%$ by blocking two of the eight openings in the first collimator to eliminate the electron elastic-scattering signal from the target. The uncertainty accounts for variation in the f_2 value between detectors and under different beam conditions.

Physics backgrounds (neutral background). A possible contribution from low-energy neutral backgrounds arising from secondary interactions of the primary scattered electrons in the collimators and magnet structure was bounded[65] to $f_3 < 0.30\%$ by subtracting f_2 from the total neutral background measured by the main detector after vetoing charged particles using thin scintillators. The asymmetry for this background was estimated[65] from a Geant4[59] simulation of the contributing processes to be $A_3 = -0.39 \pm 0.16$ p.p.m., with the dominant contribution coming from secondary interactions of electrons elastically scattered from protons.

Physics backgrounds (inelastic electrons). An unavoidable background component comes from inelastically scattered electrons that have excited the target protons to the $\Delta(1232)$ resonance, a small fraction of which enter the acceptance of the spectrometer. The fraction of the yield from inelastically scattered electrons was estimated using simulation to be $f_4 = (1.82 \pm 0.37) \times 10^{-4}$. To determine the correction to the asymmetry owing to these events, we measured the parity-violating asymmetry in the energy region near the $\Delta(1232)$ resonance during a dedicated study with a reduced spectrometer current to concentrate these electrons on the detectors. Scaling this asymmetry up to the Q^2 value of the elastic peak gives an inelastic asymmetry of $A_4 = -3.0 \pm 1.0$ p.p.m. at the elastic peak. Backgrounds from π^- particles and other hadrons were negligible.

Summary of corrections to the asymmetry and their uncertainties. The separate and combined measured asymmetries for Runs 1 and 2 are presented in Extended Data Table 2. Also shown is the breakdown of these uncertainties in terms of descending fractional significance. Extended Data Table 3 presents the numerical values of the raw asymmetry, A_{raw}, along with all systematic and acceptance correction factors used to derive the observed measured asymmetry, A_{msr}, and physics asymmetries, A_{ep}, using equations (7) and (8). Correlated uncertainties, accounted for when the two runs were combined, are also listed.

The fully corrected asymmetry is $A_{ep} = -223.5 \pm 15.0$ (statistical) ± 10.1 (systematic) p.p.b. for Run 1 and $A_{ep} = -227.2 \pm 8.3$ (statistical) ± 5.6 (systematic) p.p.b. for Run 2. The combined asymmetry is $A_{ep} = -226.5 \pm 7.3$ (statistical) ± 5.8 (systematic) p.p.b.

Data quality. Two representative tests of the consistency and the quality of the corrected asymmetries are presented below.

Null result. The Q_{weak} experiment employed signal phase locking on three independent techniques of polarized-beam helicity reversal that were used to isolate the scattering asymmetry. These were the rapid (960 Hz) reversal, the insertion of a half-wave plate in the source laser optical path at 8-h intervals and the Wien-filter reversal at monthly intervals. The half-wave plate technique is based on a mechanical action and thus is unable to induce any false asymmetries electrically or magnetically. The Wien-filter reversal rejects false asymmetries induced by beam-size (or focus) modulation. By constructing an out-of-phase, or null, asymmetry A_{null} from the latter two slow-reversal techniques, we can determine whether there are unaccounted-for false asymmetries. The full dataset-weighted null is $A_{null} = -1.75 \pm 6.51$ p.p.b., which is consistent with zero, as expected.

Asymmetry measurements. A plot of the observed main-detector asymmetry versus the Wien-filter configuration is shown in Extended Data Fig. 6. Run 1 and Run 2 were separated by a six-month accelerator shutdown, during which numerous modifications were made to the experimental apparatus and accelerator. These included upgrading the electronics associated with the beam current measurement

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

and increasing the number of associated BCMs. An electrically isolated helicity phase-locked beam-position stabilization system was enabled in the 6-MeV region of the injector during Run 2. This greatly improved the helicity-correlated stability of the beam delivered to the experiment. There were also radio-frequency-associated electronics and superconducting-cavity upgrades performed within the accelerator, unrelated to this experiment, as well as extensive upgrades to both beam polarimeters. As a consequence, the contributions of many beam-related systematic effects meaningfully changed between the two run periods. However, the resulting fully corrected physics asymmetries of the two run periods agree well. This is evidence that within the experiment's precision, the observed set of identified systematic effects is complete and their associated correction algorithms behave in a deterministic manner.

Electroweak radiative corrections and extraction of $\sin^2\theta_W$. The weak mixing angle is obtained from the proton's weak charge, taking into account energy-independent electroweak radiative corrections using[1]

$$4\sin^2\theta_W(0) = 1 - \frac{Q_w^p - \Box_{WW} - \Box_{ZZ} - \Box_{\gamma Z}(0)}{(\rho + \Delta_e)} + \Delta_e' \qquad (9)$$

Here, $\Box_{WW}$ and $\Box_{ZZ}$ are the WW and ZZ boson box-diagram radiative corrections, ρ is the renormalization of the ratio of neutral-current (Z^0) to charged-current ($W\pm$) interactions at low energy, Δ_e is the electron vertex correction to the axial Zee interaction, Δ_e' is the electron anapole moment and $\Box_{\gamma Z}(0)$ refers to the remaining energy-independent piece of the γZ-box diagram (the energy-dependent piece was discussed in 'Global fit').

The accidental suppression of the proton's weak charge in the standard model means that Q_w^p is unusually sensitive to $\sin^2\theta_W$. To see this quantitatively, our determination of Q_w^p to a precision of 6.25% results in a $\sin^2\theta_W$ precision of 0.46%. By contrast, the higher relative precision (0.59%) of the weak charge measurement[14,15] on ^{133}Cs, which is dominated by the neutron's weak charge, which is not suppressed in the standard model, leads to a $\sin^2\theta_W$ uncertainty of 0.81%, almost twice the uncertainty of our result.

The radiative corrections appearing in equation (9) are described in ref. [1] but were re-evaluated using the more recent input found in refs [2,11]. Although Q_w^p in equation (6) is defined in the Thomson limit ($Q \ll m_e$, where m_e is the electron mass) at $Q = 0$, it is determined from data in the scattering limit ($Q \gg m_e$). We chose to use radiative corrections in the Thomson limit from ref. 1 to calculate $\sin^2\theta_W(0)$ from Q_w^p. Extended Data Table 4 lists these corrections[1,2,11,67].

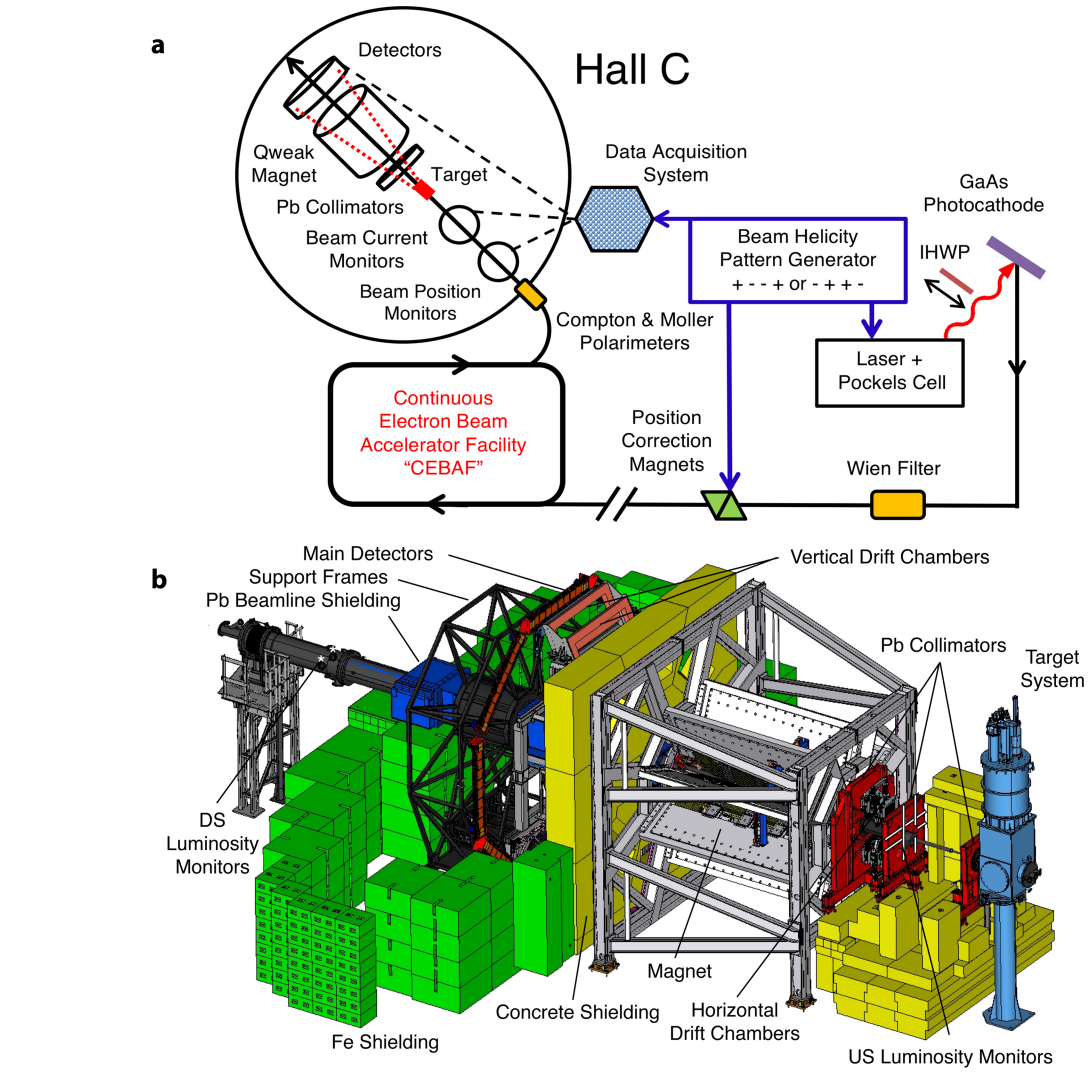
To compare our result for $\sin^2\theta_W(0)$, given by equation (9), with that of refs [2,34], we added a correction of $2\alpha/(9\pi)$, consistent with the definition of ref. [66], as discussed in ref. [20]. The result is $\sin^2\theta_W(Q=0) = 0.2384 \pm 0.0011$ in the $\overline{\text{MS}}$ scheme[2]. Shifting this to the $Q = 0.158$ GeV of the Q_{weak} experiment (a correction of -0.00012), our result is $\sin^2\theta_W(Q = 0.158\,\text{GeV}) = 0.2382 \pm 0.0111$.

Data availability. The 200 TB of raw data acquired in this work are stored at the Jefferson Laboratory data silo. The derived data supporting the findings of this study are available from the corresponding authors upon request.

Code availability. The software used for data management and analysis consisted of commercial and publicly available codes, plus experiment-specific software. Jefferson Laboratory's data management plan is available at https://scicomp.jlab.org/DataManagementPlan.pdf. The experiment-specific software is stored in a version management system (SVN & GIT) and archived at the data storage facilities of Jefferson Laboratory, in accordance with existing US regulations. Requests for this material should be addressed to the corresponding authors.

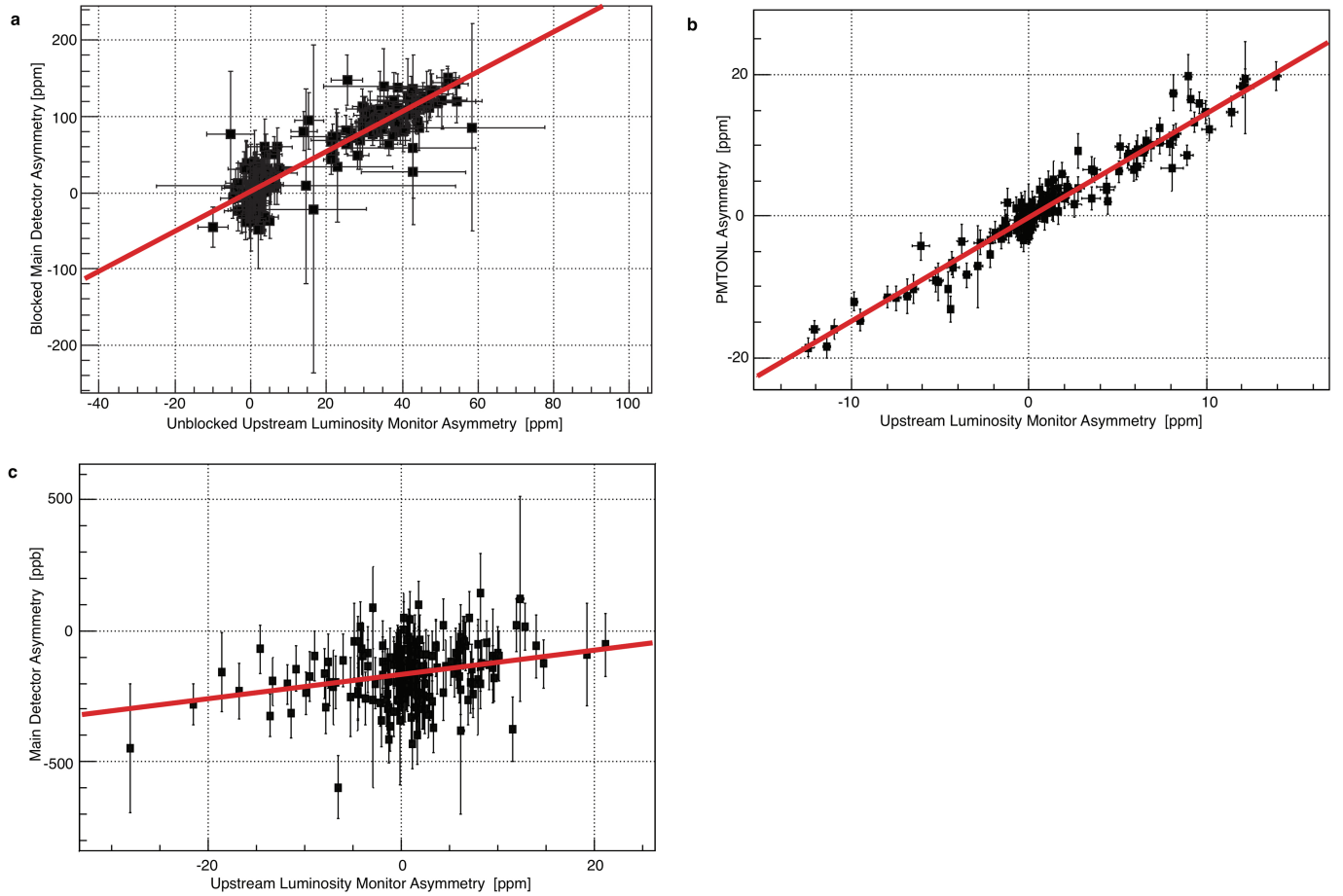
33. Musolf, M. J. et al. Intermediate-energy semileptonic probes of the hadronic neutral current. *Phys. Rep.* **239**, 1–178 (1994).
34. Erler, J. & Ramsey-Musolf, M. J. Low energy tests of the weak interaction. *Prog. Part. Nucl. Phys.* **54**, 351–442 (2005).
35. Armstrong, D. S. & McKeown, R. D. Parity-violating electron scattering and the electric and magnetic strange form factors of the proton. *Annu. Rev. Nucl. Part. Sci.* **62**, 337–359 (2012).
36. G0 Collaboration. Strange quark contributions to parity-violating asymmetries in the forward G0 electron–proton scattering experiment. *Phys. Rev. Lett.* **95**, 092001 (2005).
37. G0 Collaboration. Strange quark contributions to parity-violating asymmetries in the backward angle G0 electron scattering experiment. *Phys. Rev. Lett.* **104**, 012001 (2010).
38. HAPPEX Collaboration. Parity-violating electroweak asymmetry in *ep* scattering. *Phys. Rev. C* **69**, 065501 (2004).
39. HAPPEX Collaboration. Constraints on the nucleon strange form-factors at $Q^2 \sim 0.1$ GeV^2. *Phys. Lett. B* **635**, 275–279 (2006).
40. HAPPEX Collaboration. Precision measurements of the nucleon strange form factors at $Q^2 \sim 0.1$ GeV^2. *Phys. Rev. Lett.* **98**, 032301 (2007).
41. HAPPEX Collaboration. New precision limit on the strange vector form factors of the proton. *Phys. Rev. Lett.* **108**, 102001 (2012).
42. Spayde, D. T. et al. The strange quark contribution to the proton's magnetic moment. *Phys. Lett. B* **583**, 79–86 (2004).
43. Maas, F. E. et al. Measurement of strange quark contributions to the nucleon's form-factors at $Q^2 = 0.230$ (GeV/c)^2. *Phys. Rev. Lett.* **93**, 022002 (2004).
44. Maas, F. E. et al. Evidence for strange quark contributions to the nucleon's form-factors at $Q^2 = 0.108$ (GeV/c)^2. *Phys. Rev. Lett.* **94**, 152001 (2005).
45. Baunack, S. et al. Measurement of strange quark contributions to the vector form factors of the proton at $Q^2 = 0.22$ (GeV/c)^2. *Phys. Rev. Lett.* **102**, 151803 (2009).
46. HAPPEX Collaboration. Parity-violating electron scattering from ^{4}He and the strange electric form factor of the nucleon. *Phys. Rev. Lett.* **96**, 022003 (2006).
47. Balaguer Ríos, D. et al. Measurement of the parity violating asymmetry in the quasielastic electron-deuteron scattering and improved determination of the magnetic strange form factor and the isovector anapole radiative correction. *Phys. Rev. D* **94**, 051101 (2016).
48. SAMPLE Collaboration. Parity violating electron deuteron scattering and the proton's neutral weak axial vector form-factor. *Phys. Rev. Lett.* **92**, 102003 (2004).
49. González-Jiménez, R., Caballero, J. A. & Donnelly, T. W. Global analysis of parity-violating asymmetry data for elastic electron scattering. *Phys. Rev. D* **90**, 033002 (2014).
50. Zhu, S., Puglia, S. J., Holstein, B. R. & Ramsey-Musolf, M. J. Nucleon anapole moment and parity-violating *ep* scattering. *Phys. Rev. D* **62**, 033008 (2000).
51. Kelly, J. J. Simple parametrization of nucleon form factors. *Phys. Rev. C* **70**, 068202 (2004).
52. Galster, S., Klein, H., Moritz, J., Schmidt, K. H. & Wegener, D. Elastic electron–deuteron scattering and the electric neutron form factor at four momentum transfers 5 fm$^{-2} < q^2 < 14$ fm^{-2}. *Nucl. Phys. B* **32**, 221–237 (1971).
53. Venkat, S., Arrington, J., Miller, G. A. & Zhan, X. Realistic transverse images of the proton charge and magnetic densities. *Phys. Rev. C* **83**, 015203 (2011).
54. Rislow, B. C. & Carlson, C. E. Modification of electromagnetic structure functions for the γZ -box diagram. *Phys. Rev. D* **88**, 013018 (2013).
55. Grames, J. et al. Two Wien filter spin flipper. In *Proc. 2011 Particle Accelerator Conference* (eds Satogata, T. & Brown, K.) 862–864 (IEEE, New York, 2011).
56. Arrington, J., Blunden, P. G. & Melnitchouk, W. Review of two-photon exchange in electron scattering. *Prog. Part. Nucl. Phys.* **66**, 782–833 (2011).
57. Qweak Collaboration. Beam normal single spin asymmetry measurements from Q_{weak}. Preprint at https://arxiv.org/abs/1604.04602 (2016).
58. Kargiantoulakis, E. *A Precision Test of the Standard Model via Parity-Violating Electron Scattering in the Q_{weak} Experiment*. PhD thesis, Univ. Virginia (2015); https://misportal.jlab.org/ul/publications/view_pub.cfm?pub_id=14261.
59. Agostinelli, S. et al. Geant4 – a simulation toolkit. *Nucl. Instrum. Meth. A* **506**, 250–303 (2003).
60. Abrahamyan, S. et al. New measurements of the transverse beam asymmetry for elastic electron scattering from selected nuclei. *Phys. Rev. Lett.* **109**, 192501 (2012).
61. Gorchtein, M. & Horowitz, C. J. Analyzing power in elastic scattering of the electrons off a spin-0 target. *Phys. Rev. C* **77**, 044606 (2008).
62. Hauger, M. et al. A high-precision polarimeter. *Nucl. Instrum. Meth. A* **462**, 382–392 (2001).
63. Magee, J. et al. A novel comparison of Møller and Compton electron-beam polarimeters. *Phys. Lett. B* **766**, 339–344 (2017).
64. Horowitz, C. J. Parity violating elastic electron scattering from ^{27}Al and the QWEAK measurement. *Phys. Rev. C* **89**, 045503 (2014).
65. McHugh, M. *A Measurement of the Transverse Asymmetry in Forward-Angle Electron-Carbon Scattering Using the Q_{weak}Apparatus*. PhD thesis, George Washington Univ. (2017); https://misportal.jlab.org/ul/publications/view_pub.cfm?pub_id=14918.
66. Ferroglia, A., Ossola, G. & Sirlin, A. Bounds on M_W, M_t, $\sin^2\theta_{\text{eff}}^{\text{lept}}$. *Eur. Phys. J. C* **35**, 501–507 (2004).
67. Bethke, S. αs 2002. *Nucl. Phys. B Proc. Sup.* **121**, 74–81 (2003).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



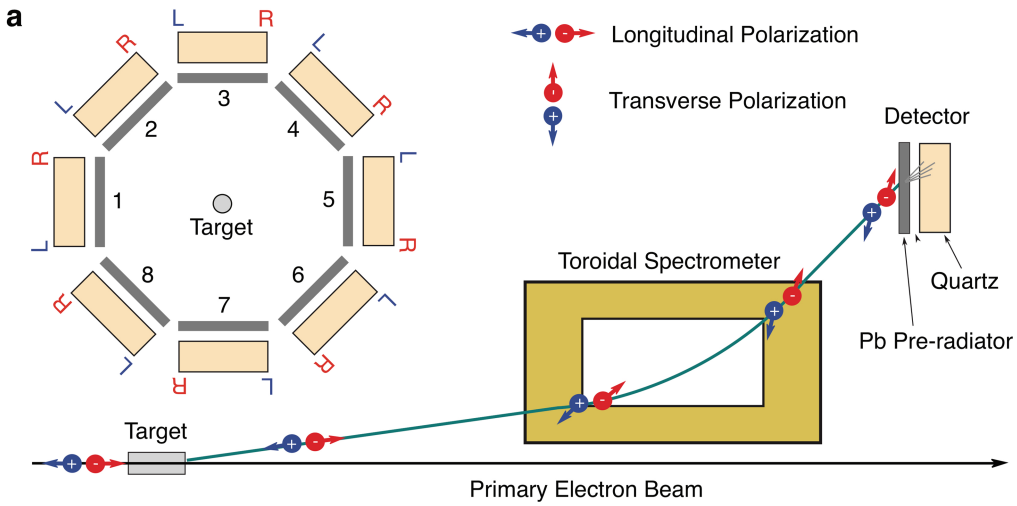
Extended Data Fig. 1 | Apparatus. a, Schematic of critical accelerator components and the Q_{weak} apparatus[4]. The electron beam is generated at the photocathode, accelerated by the Continuous Electron Beam Accelerator Facility (CEBAF) and sent to experimental Hall C, where it is monitored by beam position monitors and beam current monitors. The insertable half-wave plate (IHWP) provides slow reversal of the electron beam helicity. The data acquisition system records the data. **b**, Computer-aided design drawing of the experimental apparatus. **c**, The Q_{weak} apparatus, before the final shielding configuration was installed. **d**, Interior of the hut shielding the detectors, showing two of the Cherenkov detectors (right) and a pair of tracking chambers (left).

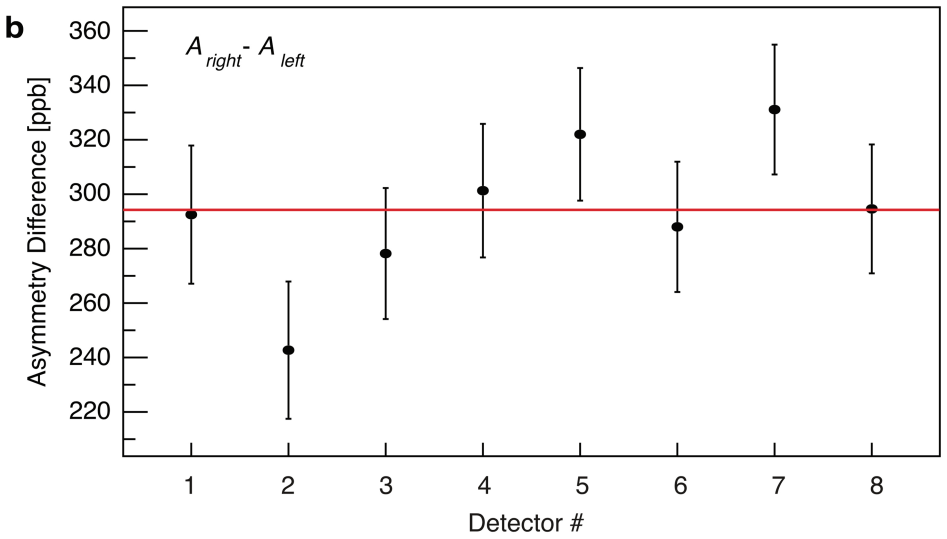
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

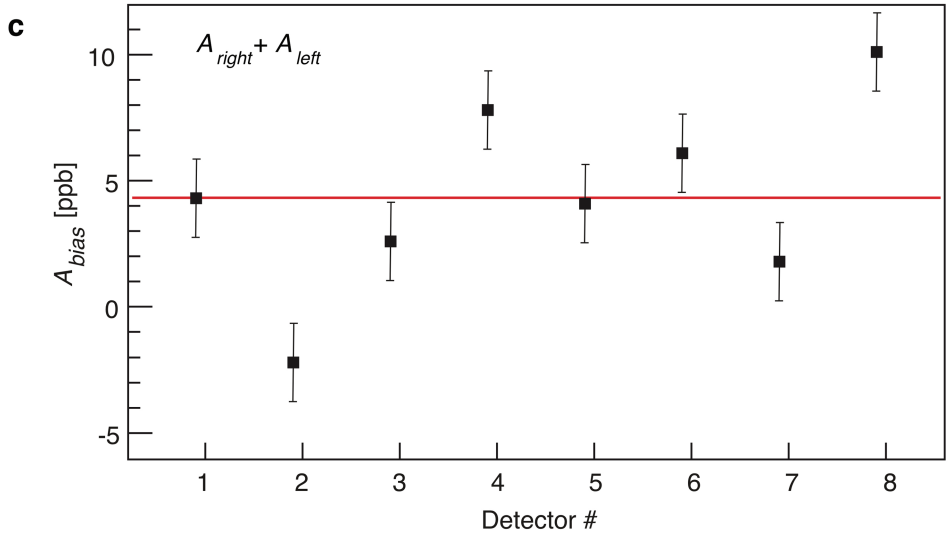


Extended Data Fig. 2 | Beamline background. Determination of A_{bb}, the false asymmetry arising from beamline background events. Uncertainties are 1 s.d. **a**, Correlation of the main detector asymmetry to that of the upstream luminosity monitors, measured when the signal from elastically scattered electrons in the main detectors was blocked at the first collimator. **b**, Correlation of asymmetries from the upstream luminosity monitors with one of the other background detectors (a bare PMT located in the detector shield house). **c**, Correlation of the unblocked main detector asymmetry to that of the upstream luminosity monitor for Run 2. Our A_{bb} determination was based on this slope.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

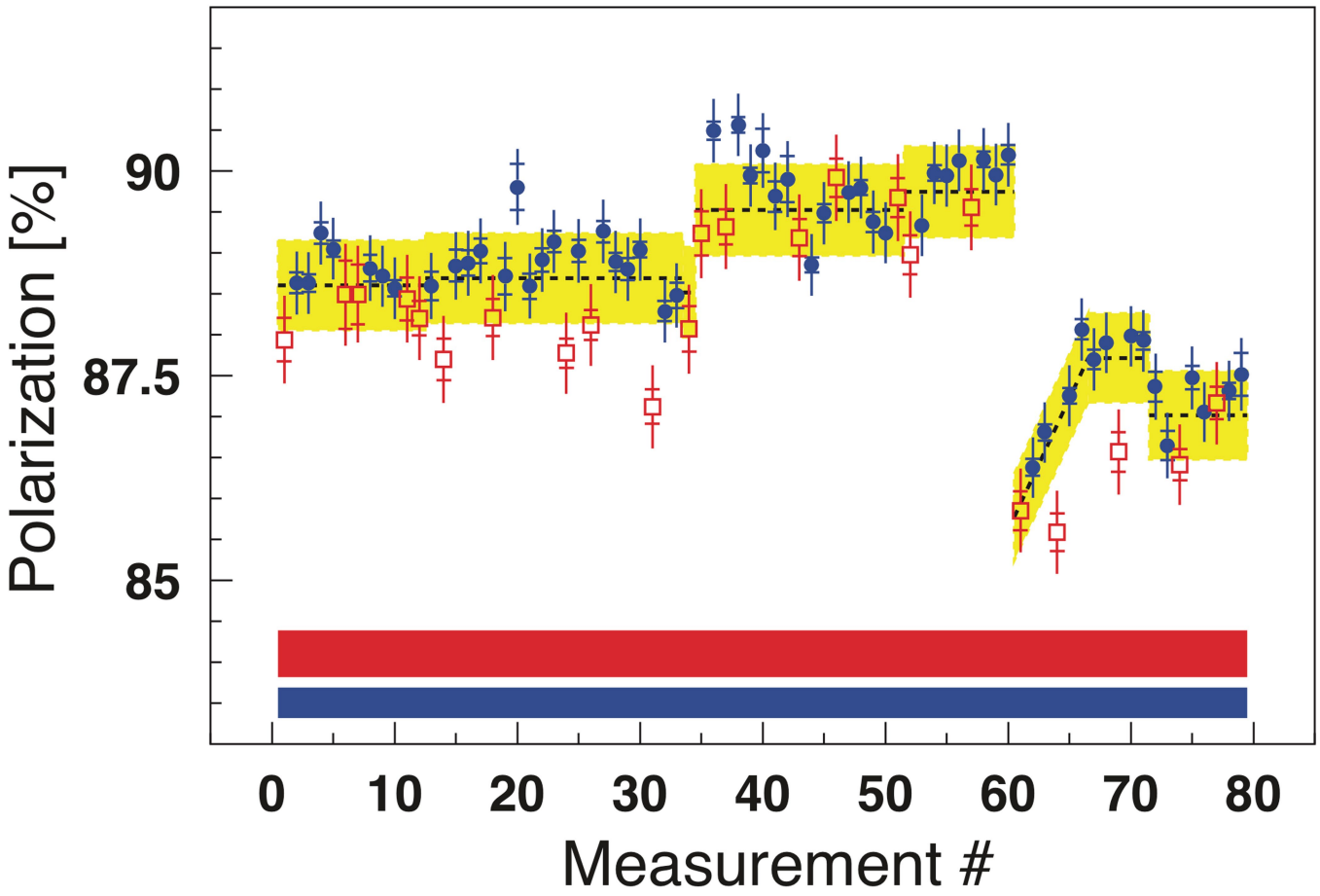






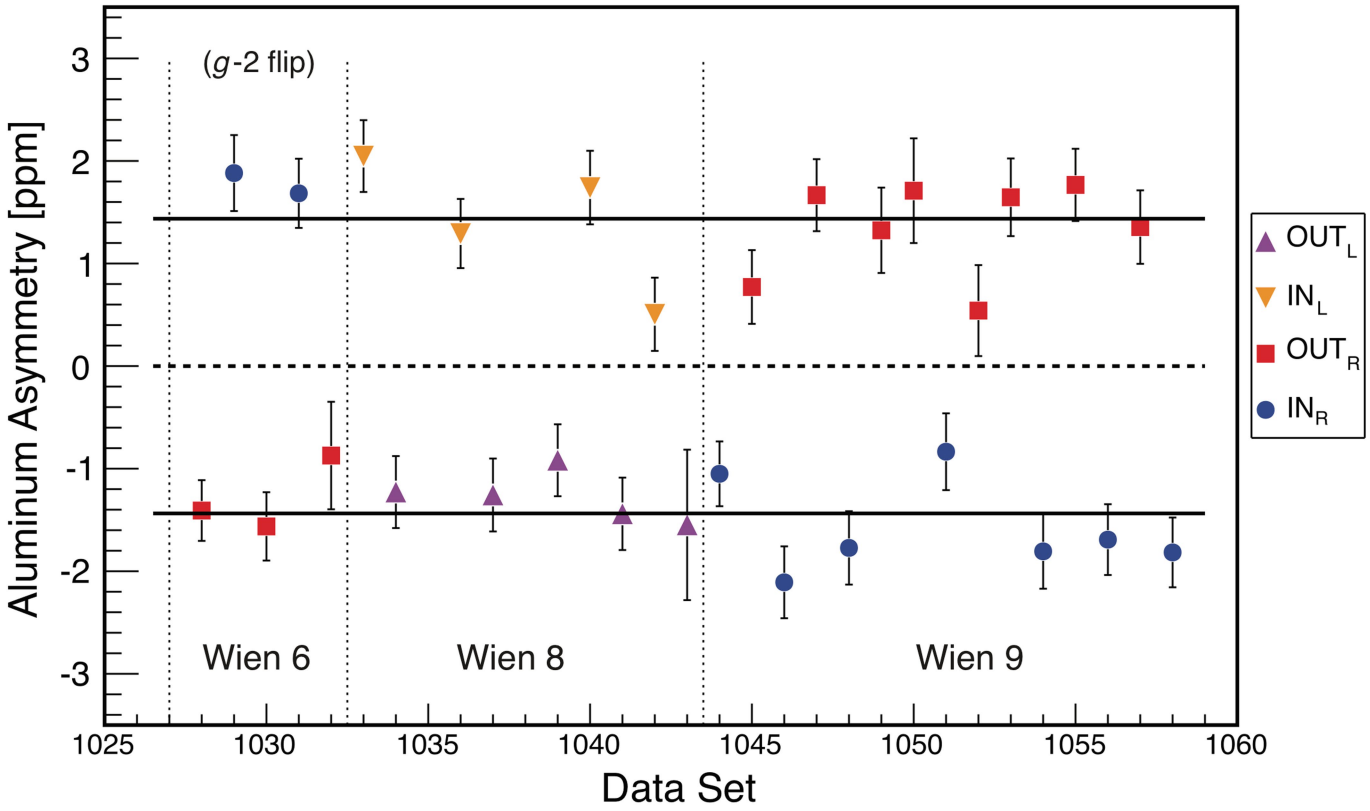
Extended Data Fig. 3 | Rescattering bias. a, Schematic illustrating the precession of longitudinally polarized electrons through the spectrometer magnet, generating sizeable transverse spin components upon arrival at the detector array (spin directions indicated by red and blue arrows for the two electron helicity states). An end-view of the detector array, indicating the right (R) and left (L) PMT positions, is shown on the left. **b**, Difference between the asymmetry measured by the two (R and L) PMT tubes versus the detector number (Run 2 data). **c**, Calculated rescattering bias A_{bias} versus detector number, with the eight-detector-averaged value shown by the red lines. Uncertainties (1 s.d.) are systematic.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



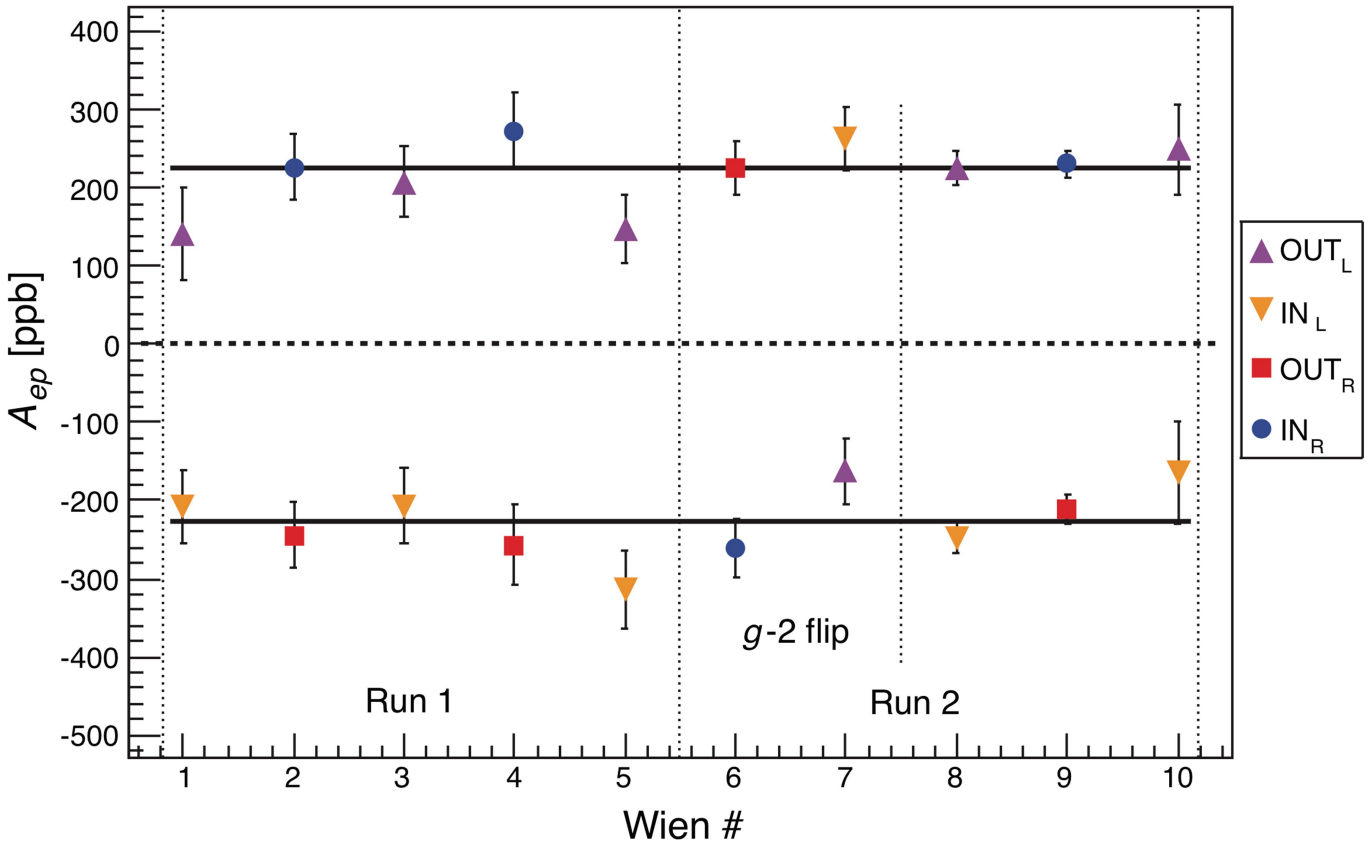
Extended Data Fig. 4 | Electron beam polarization. Measurements from the Compton (closed blue circles) and Møller (open red squares) polarimeters during Run 2. Inner error bars denote statistical uncertainties and outer error bars show the statistical and point-to-point systematic uncertainties added in quadrature. Normalization, or scale-type, uncertainties are shown by the solid blue (Compton) and red (Møller) bands. All uncertainties are 1 s.d. The yellow band shows the derived polarization values used in the evaluation of the parity-violating asymmetry A_{ep}. The time dependence of the reported polarization is driven primarily by the continuous Compton measurements, with a small-scale correction (0.21%, not included in this figure) determined from an uncertainty-weighted global comparison of the Compton and Møller polarimeters.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 5 | Asymmetry from aluminium. Parity-violating asymmetry from the aluminium alloy target versus the dataset number. All uncertainties are 1 s.d. The labels 'IN' and 'OUT' refer to the state of the insertable half-wave plate at the electron source, which generated a 180° flip of the electron spin when IN. The subscripts denote the setting of the Wien filter, with L and R corresponding to the presence and absence, respectively, of an additional 180° rotation of the spin direction of the electron beam. A period in which a further 180° flip was generated through ($g_e - 2$) precession (g_e, electron gyromagnetic ratio) via a modified accelerator configuration during Wien 6 is indicated. The combinations OUT–R and IN–L with no ($g_e - 2$) spin flip reveal the physical sign of the asymmetry. Solid lines represent the time-averaged values, and the horizontal dashed line indicates zero asymmetry. The vertical dashed lines delineate particular data subsets with a given Wien filter setting.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Asymmetry from the proton. Observed parity-violating asymmetry A_{ep} after all corrections, versus the dataset number (acquired in the double-Wien-filter configuration). The Wien filter reversed the beam helicity at approximately monthly intervals. The subscripts denote the setting of the Wien filter as L or R, corresponding to the presence or absence, respectively, of a 180° rotation of the spin direction of the electron beam. IN and OUT refer to the state of the insertable half-wave plate at the electron source, generating an additional 180° flip of the spin when IN. A period in which a further 180° flip was generated through $(g_e - 2)$ precession via a modified accelerator configuration is indicated. The combinations OUT–R and IN–L with no $(g_e - 2)$ flip reveal the physical sign of the asymmetry. Solid lines represent the time-averaged values and the dashed line indicates zero asymmetry. The uncertainties (1 s.d.) shown are those of the corresponding A_{msr} values (see text) only—that is, they do not include time-independent uncertainties—so as to illustrate the time stability of the results. The weighted mean and P-value of the upper OUT–L and IN–R data are 226.9 ± 10.2, $P = 0.59$ (upper solid line), respectively. For the opposite combination, OUT–R and IN–L, we find a weighted mean of -226.1 ± 10.5 and $P = 0.36$ (lower solid line).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Helicity-correlated beam parameter differences and sensitivities

Beam Parameter	Run 1 $\Delta\chi_i$	Run 2 $\Delta\chi_i$	Typical $\partial A/\partial\chi_i$
X	-3.5 ± 0.1 nm	-2.3 ± 0.1 nm	-2 ppb/nm
X'	-0.30 ± 0.01 nrad	-0.07 ± 0.01 nrad	50 ppb/nrad
Y	-7.5 ± 0.1 nm	0.8 ± 0.1 nm	< 0.2 ppb/nm
Y'	-0.07 ± 0.01 nrad	-0.04 ± 0.01 nrad	< 3 ppb/nrad
Energy	-1.69 ± 0.01 ppb	-0.12 ± 0.01 ppb	-6 ppb/ppb

The beam parameter differences and typical detector sensitivities for the five measured beam parameters for Run 1 and Run 2. Uncertainties are 1 s.d.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Asymmetries and their corrections

Period	Asymmetry (ppb)	Stat. Unc. (ppb)	Syst. Unc. (ppb)	Tot. Uncertainty (ppb)
Run 1	-223.5	15.0	10.1	18.0
Run 2	-227.2	8.3	5.6	10.0
Run 1 and 2 combined with correlations	-226.5	7.3	5.8	9.3

Quantity	Run 1 error (ppb)	Run 1 fractional	Run 2 error (ppb)	Run 2 fractional
BCM Normalization: A_{BCM}	5.1	25%	2.3	17%
Beamline Background: A_{BB}	5.1	25%	1.2	5%
Beam Asymmetries: A_{beam}	4.7	22%	1.2	5%
Rescattering bias: A_{bias}	3.4	11%	3.4	37%
Beam Polarization: P	2.2	5%	1.2	4%
Target windows: A_1	1.9	4%	1.9	12%
Kinematics: R_{Q^2}	1.2	2%	1.3	5%
Total of others	2.5	6%	2.2	15%
Combined in quadrature	10.1		5.6	

Top, corrected asymmetries A_{ep} for the Run 1 and Run 2 datasets, and the combined value, with their statistical (Stat. Unc.), systematic (Syst. Unc.) and total (Tot.) uncertainties (1 s.d.), in parts per billion (ppb). Bottom, fractional quadrature contributions $(\sigma_i/\sigma_{\mathrm{tot}})^2$ to the systematic uncertainty (1 s.d.) on A_{ep} for Run 1 and Run 2. Only error sources with fractional contribution $\geq 5\%$ in one of the runs are shown.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 3 | Raw asymmetries and their corrections

Quantity	Run 1	Run 2	Correlation
A_{raw}	-192.7 ± 13.2 ppb	-170.7 ± 7.3 ppb	—
A_{T}	0 ± 1.1 ppb	0 ± 0.7 ppb	0
A_{L}	1.3 ± 1.0 ppb	1.2 ± 0.9 ppb	1
A_{BCM}	0 ± 4.4 ppb	0 ± 2.1 ppb	0.67
A_{BB}	3.9 ± 4.5 ppb	-2.4 ± 1.1 ppb	0
A_{beam}	18.5 ± 4.1 ppb	0.0 ± 1.1 ppb	0
A_{bias}	4.3 ± 3.0 ppb	4.3 ± 3.0 ppb	1
A_{msr}	-164.6 ± 15.5 ppb	-167.5 ± 8.4 ppb	—
P	87.66 ± 1.05 %	88.71 ± 0.55 %	0.19
f_1	2.471 ± 0.056 %	2.516 ± 0.059 %	1
A_1	1.514 ± 0.077 ppm	1.515 ± 0.077 ppm	1
f_2	0.193 ± 0.064 %	0.193 ± 0.064 %	1
f_3	0.12 ± 0.20 %	0.06 ± 0.12 %	1
A_3	-0.39 ± 0.16 ppm	-0.39 ± 0.16ppm	1
f_4	0.018 ± 0.004 %	0.018 ± 0.004 %	1
A_4	-3.0 ± 1.0 ppm	-3.0 ± 1.0ppm	1
R_{RC}	1.010 ± 0.005	1.010 ± 0.005	1
R_{Det}	0.9895 ± 0.0021	0.9895 ± 0.0021	1
R_{Acc}	0.977 ± 0.002	0.977 ± 0.002	1
R_{Q^2}	0.9928 ± 0.0055	1.0 ± 0.0055	1
R_{tot}	0.9693 ± 0.0080	0.9764 ± 0.0080	1
$\sum f_i$	2.80 ± 0.22 %	2.78 ± 0.15 %	1

The raw measured asymmetries A_{raw} for the two run periods, and all the corrections (including for false asymmetries, backgrounds, beam polarization and detector acceptance) that were applied to extract the final asymmetry A_{ep} from A_{raw} (see text). The f_i values are dilutions to the signal, A_i are false or background process asymmetries, P is the beam polarization and R_i are multiplicative factors. The net multiplicative correction R_{tot} and the total dilution are also indicated, as well as the values of A_{msr}, the asymmetry after the corrections for the false asymmetries (see equation (7)). The correlations used to combine the two runs are provided in the final column. Uncertainties are 1 s.d.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 4 | Radiative corrections

Term	Expression	Value	Reference
ρ_{NC}	$1+\Delta_\rho$	1.00066	1, 2
Δ_e	$-\alpha/2\pi$	−0.001161	1, 2
Δ'_e	$-\frac{\alpha}{3\pi}(1-4\hat{s}^2)\left[\ln\left(\frac{M_Z^2}{m_e^2}\right)+\frac{1}{6}\right]$	−0.001411	1, 2
$\hat{\alpha}$	$\equiv \alpha(M_Z)$	1/127.95	1, 2
$\hat{s}^2$	$=1-\hat{c}^2 \equiv \sin^2\theta_W(M_Z)$	0.23129	1, 2
$\alpha_s(M_W^2)$	-	0.12072	67
$\square_{WW}$	$\frac{\hat{\alpha}}{4\pi\hat{s}^2}\left[2+5\left(1-\frac{\alpha_s(M_W^2)}{\pi}\right)\right]$	0.01831	1, 2
$\square_{ZZ}$	$\frac{\hat{\alpha}}{4\pi\hat{s}^2\hat{c}^2}\left[9/4-5\hat{s}^2\right]\left(1-4\hat{s}^2+8\hat{s}^2\right)\left(1-\frac{\alpha_s(M_Z^2)}{\pi}\right)$	0.00185	1, 2
$\square_{\gamma Z}$	axial-vector hadron piece of $\square_{\gamma Z}$: $\Re\mathrm{e}\,\square^{\mathrm{A}}_{\gamma\mathrm{Z}}$	0.0044	11

Numerical values used for the electroweak radiative corrections in equation (9).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0085-3

Challenging local realism with human choices

The BIG Bell Test Collaboration*

A Bell test is a randomized trial that compares experimental observations against the philosophical worldview of local realism[1], in which the properties of the physical world are independent of our observation of them and no signal travels faster than light. A Bell test requires spatially distributed entanglement, fast and high-efficiency detection and unpredictable measurement settings[2,3]. Although technology can satisfy the first two of these requirements[4–7], the use of physical devices to choose settings in a Bell test involves making assumptions about the physics that one aims to test. Bell himself noted this weakness in using physical setting choices and argued that human 'free will' could be used rigorously to ensure unpredictability in Bell tests[8]. Here we report a set of local-realism tests using human choices, which avoids assumptions about predictability in physics. We recruited about 100,000 human participants to play an online video game that incentivizes fast, sustained input of unpredictable selections and illustrates Bell-test methodology[9]. The participants generated 97,347,490 binary choices, which were directed via a scalable web platform to 12 laboratories on five continents, where 13 experiments tested local realism using photons[5,6], single atoms[7], atomic ensembles[10] and superconducting devices[11]. Over a 12-hour period on 30 November 2016, participants worldwide provided a sustained data flow of over 1,000 bits per second to the experiments, which used different human-generated data to choose each measurement setting. The observed correlations strongly contradict local realism and other realistic positions in bipartite and tripartite[12] scenarios. Project outcomes include closing the 'freedom-of-choice loophole' (the possibility that the setting choices are influenced by 'hidden variables' to correlate with the particle properties[13]), the utilization of video-game methods[14] for rapid collection of human-generated randomness, and the use of networking techniques for global participation in experimental science.

Bell tests, like Darwin's studies of finches and Galileo's observations of the moons of Jupiter, use empirical methods to address questions previously accessible only by other means, for example, by philosophy or theology[15]. Local realism—that is, realism plus relativistic limits on causation—was debated by Einstein and Bohr using metaphysical arguments and recently has been rejected by Bell tests[4–7] that closed all technical 'loopholes'. For example, the 'detection-efficiency loophole' describes the possibility that the observed statistics are inaccurate owing to selection bias, and is closed by using high-efficiency detection and statistical methods that include all trials in the analysis. Recent work on device-independent quantum information[16] shows how Bell inequality violation (BIV) can also challenge causal determinism[17], another topic that was formerly accessible only by metaphysics[18]. Central to both applications is the use of free variables to choose measurements: in the words of Aaronson[19], "Assuming no preferred reference frames or closed timelike curves, if Alice and Bob have genuine 'freedom' in deciding how to measure entangled particles, then the particles must also have 'freedom' in deciding how to respond to the measurements".

Previous Bell tests used physical devices[20,21] to 'decide' for Alice and Bob, and thus demonstrated only a relation among physical processes: if some processes are 'free' in the required sense (see Methods, 'Freedom' in Bell tests), then other processes are similarly 'free'. In the language of strong Bell tests, this conditional relation leaves open the freedom-of-choice loophole (FOCL), which describes the possibility that 'hidden variables' influence the setting choices. Because we cannot guarantee such freedom within local realism, the tests must assume physical indeterminacy in the hidden-variable theory[2]. Laboratory methods can tighten, but never close, this loophole[2–6].

Gallicchio, Friedman and Kaiser[22] have proposed choosing settings by observing cosmic sources at the edge of the visible Universe. A BIV under such conditions could only be explained within local realism if events across history conspired to produce the measured outcomes[23,24]. Bell himself argued that human choices could be considered 'free variables' in a Bell test[8] (see Methods, John Stewart Bell on 'free variables'), and noted the impracticality of using humans with 1970s' technologies. Here we implement Bell's idea, using modern crowd-sourcing, networking and gamification[14] techniques. In this BIG Bell Test (BBT), Alice and Bob of Aaronson's formulation are real people. Assuming no faster-than-light communication, such experiments can prove the conditional premise that if human will is free, there are physical events (the measurement outcomes in the Bell tests) that are intrinsically random, that is, impossible to predict[25]. We note that although this argument in no way uses the theory of quantum mechanics, it arrives at one of the theory's most profound claims. Intrinsic randomness supported by a BIV is central to device-independent quantum technologies[16,26].

It is perhaps surprising that human choices, which are known to contain statistical regularities[27], are suitably unpredictable for a Bell test. Recent works on the statistical analysis of Bell tests[3,28,29] show that sequence randomness—that is, the absence of patterns and correlations in the sequence of choices—is not, per se, a requirement for the rejection of local realism. Rather, statistical independence of choices from the hidden variables that describe possible measurement outcomes is required (see Methods, 'Freedom' in Bell tests). This independence can fail in different ways, which are categorized into named loopholes. The 'locality loophole' describes the possibility that a choice at one station could influence a measurement result at the other station. The term 'locality' reflects one way of blocking this possibility, by space-like separation of the choice and measurement events (see Methods, Use of 'freedom-of-choice loophole' and 'locality loophole' in this work).

Patterns strongly affect statistical strength in experiments that aim to close the locality loophole by space-like separation—they allow current choices to be predicted from earlier choices, which have had more time to reach the distant measurement. As described below, the BBT tightens the locality loophole by using many independent experiments instead of space-like separation. Furthermore, the human capacity for free choice removes the need for assumptions about physical indeterminism, allowing the FOCL to be closed. Thus, although human choices show imperfect sequence randomness, they nonetheless enable a strong rejection of local realism with the BBT strategy.

A major obstacle to a Bell test with humans has been the difficulty in generating enough choices for a statistically significant test. A person can generate roughly three random bits per second, while a strong test may require millions of setting choices in a time span of minutes to hours, depending on the speed and stability of the experiment. To achieve such rates, we crowd-sourced the basis choices, recruiting about

*A list of authors and their affiliations appears at the end of the paper.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

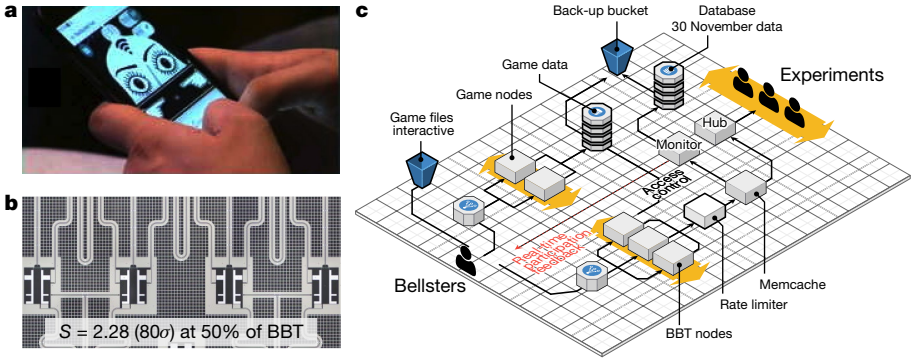


Fig. 1 | Structure of the BBT. a, Human participants, or Bellsters, enter 0s and 1s in an online video game that incentivizes sustained generation of unpredictable input. Image: ICFO/Maria Pascual (Kaitos Games). **b**, Experiments use Bellster-generated bits to control measurement-defining elements, such as wave-plates for photons or microwave pulses for matter qubits. Shown is a micrograph of superconducting qubits used in experiment (7), with the measured Clauser–Horne–Shimony–Holt Bell parameter S mid-way through the BBT. **c**, A cloud-based networking system integrates the activities shown in **a** and **b**, serving game elements to Bellsters, distributing input bits to connected laboratories and providing in-game feedback about the experimental use of the player's input. Through this system, Bellsters are given direct—although brief—control of the experimental apparatus, so that each measurement setting is determined by a single human choice, which is traceable to a given user and time of entry (see Methods, Networking strategy and architecture).

100,000 participants, the 'Bellsters', over the course of the project. Each choice made by a participant, encoded as a bit (0 or 1), was entered in an internet-connected device, such as the participant's mobile phone. Servers relayed these bits to the 13 experiments (see Fig. 1), which used them to determine individual settings without re-use (except experiment (2)). To encourage participants to contribute a larger number of more unpredictable bits, the input was collected in the context of a video game, The BIG Bell Quest (available at https://museum.thebigbelltest.org/quest/), implemented in JavaScript to run directly in a device's web browser.

The BIG Bell Quest was designed to reward sustained, high-rate input of unpredictable bits, while being engaging and informative (see Methods, Gamification). An interactive explanation first describes quantum nonlocality and the role played by participants and experimenters in the BBT. The player is then tasked with entering a given number of unpredictable bits within a limited time. A machine learning algorithm (MLA) attempts to predict each input bit by modelling the user's input as a Markov process and updating the model parameters using reinforcement learning (see Methods, Prediction engine). Scoring and level completion reflect the degree to which the MLA predicts the player's input, motivating players to consider their own predictability and take conscious steps to reduce it. However, the MLA does not act as a filter, and all input is passed to the experiments. Bellsters' input showed unsurprising deviations from ideal randomness[27]; for example, $P(0) \approx 0.5237$ (bias towards 0), whereas adjacent bits show $P(01)+P(10) \approx 0.6406$ (excess of alternation).

Modern video-game elements were incorporated to boost engagement (animation and sound), encourage persistent play (progressive levels, 'power-ups', 'boss battles' and leaderboards) and to recruit new players (group formation and postings to social networks). Different level scenarios illustrate the key elements of the BBT (human input, global networking and measurements on quantum systems), and boss battles against the Oracle (see Methods) convey the conceptual challenge of unpredictability. Level completion is rewarded with (i) a report on how many bits from that level were used in each experiment running at that time, (ii) a 'curious fact' about statistics, Bell tests or the various experiments and, if the participant is lucky, (iii) one of several videos recorded in the participating laboratories, explaining the experiments. The game and BBT website are available in Chinese, English, Spanish, French, German, Italian and Catalan, making them accessible to roughly three billion speakers of these languages.

To synchronize participant activity with experimental operation, the Bell tests were scheduled to take place on a single day, Wednesday 30 November 2016. The date was chosen so that most schools worldwide would be in session and to avoid competing media events, such as the US presidential election. Participants were recruited through a variety of channels, including traditional and social media, as well as school and science museum outreach programmes, with each partner institution handling recruitment in their familiar geographical regions and languages. The media campaign focused on the nature of the experiment and the need for human participants. The press often communicated this with headlines such as "Quantum theory needs your help" (China Daily). A first, small campaign in early October 2016 began spreading the story by word of mouth and a second, large campaign on 29–30 November of that year was made to attract a wide participant base. The media campaign generated at least 230 headlines in the printed and online press, radio and television.

The data networking architecture of the BBT, shown in Fig. 1, includes elements of instant messaging and online gaming and is designed to efficiently serve a fluctuating number of simultaneous users that is not known in advance and could range from 10 to 100,000. A gaming component handles the BBT website, participant account management, delivery of the game code (JavaScript and video), score records and leaderboards. In parallel, a messaging component handles data conditioning, streaming to experiments and reporting of participant choices generated via the game. Horizontal scaling is used in both components; participants do not connect to the servers directly, but connect to dynamic load balancers that spread the input among a pool of servers that are dynamically scaled in response to the load. The timing of input bits (but not their values) was used to identify robot participants and remove their input from the data stream, although game operation was unchanged to avoid alerting the robots' masters. A single, laboratory-side server received data from the participant-side servers, concatenated the user input and streamed it to the laboratories at laboratory-defined rates (see Methods, Networking strategy and architecture).

According to global time zoning, 30 November defines a 51-h window, from 0:00 UTC (coordinated universal time) plus 14 h (for example, Samoa) to 23:59 UTC minus 12 h (for example, Midway Island). Nevertheless, most participants contributed during a 24-h window centred on 18:00 UTC. The recruitment of participants was geographically uneven, with a notable failure to recruit large numbers of participants from Africa. Despite this, the latitude zones of Asia–Oceania, Europe–Africa and the Americas had comparable participation, which proved important for the experiment. As shown in Fig. 2, input from any single region dropped to low values during the local early morning but was compensated by high input from other regions, resulting in a sustained high global bitrate. Over the 12-h period from 09:00 UTC to 21:00 UTC on 30 November 2016, the input exceeded 10^3 bits per second, allowing the majority of the experiments to run at their full speeds. Several

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

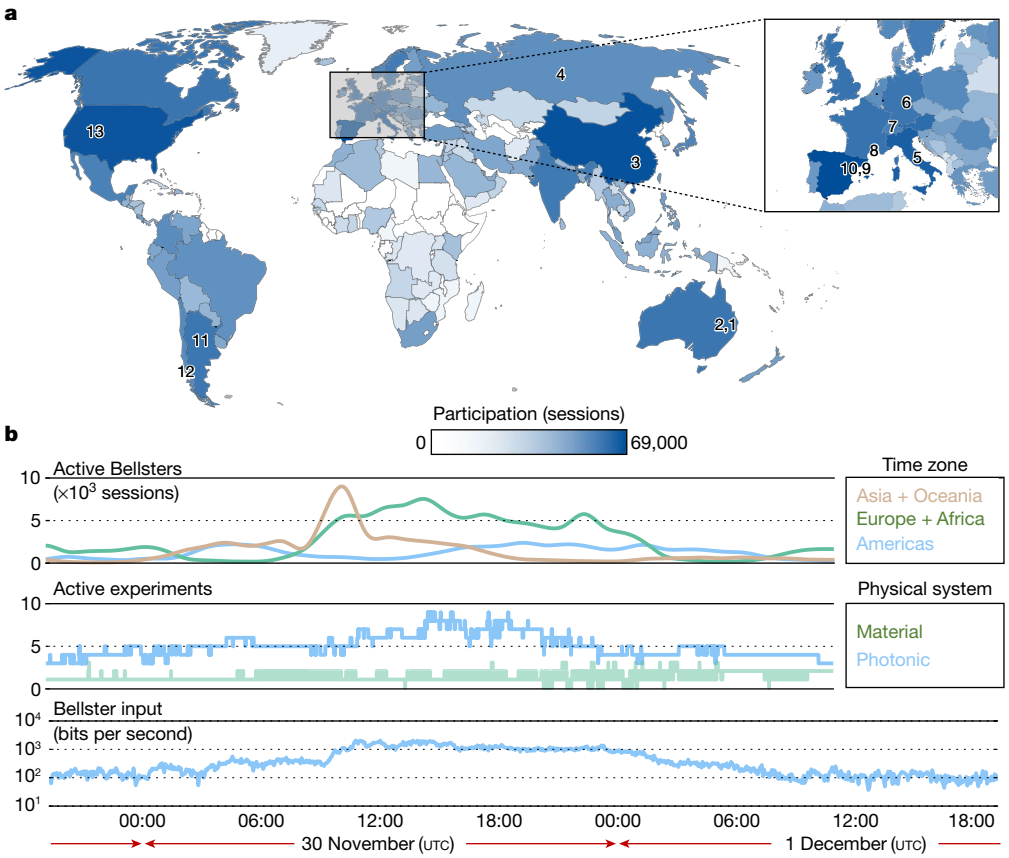


Fig. 2 | Geography and timing of the BBT. a, Locations of the 13 BBT experiments, ordered from east to west. The index numbers label the experiments, which are summarized in Table 1. Shading shows total sessions by country. Eight sessions from Antarctica are not shown. Map created by G. Colangelo using data from OpenStreetMaps, rendered in Wolfram Mathematica. **b**, Temporal evolution of the project. The top graph shows the number of live sessions versus time for different-continent groups, which exhibits a large drop in the local early morning in each region. The spike in the participation of the Asian group around 11:00 UTC coincides with a live-streamed event in Barcelona, hosted by D. Jiménez and the CosmoCaixa science museum, re-broadcast live in Chinese by L.-F. Yuan and the University of Science and Technology of China (USTC). The middle graph shows the number of connected laboratories versus time, divided into experiments using only photons and experiments with at least one material component (such as atoms or superconductors). The bottom graph shows the input bitrate versus time. The data flow remains nearly constant despite regional variations, with Asian Bellsters handing off to Bellsters from the Americas in the critical period 12:00–00:00 UTC. Session data from Google Analytics.

experiments posted their results live on social networks. Owing to their high speeds, experiments (12) and (13) accumulated participants' input to use in short bursts. As determined by separate tests, for example of interference visibility, experiments (9) and (12) were not in condition to observe a BIV on the day of the event; they reported later results using stored input.

Because the Earth has a diameter of only 43 light milliseconds, human choices are too slow to be space-like separated from the measurements. This leaves open the locality loophole regarding the influence of choices on remote detection. Influence of Alice's measurement setting on Bob's detection (and vice versa) is nonetheless excluded by space-like separation in experiments (3) and (13) (see Supplementary Information). To tighten the locality loophole, we employ a strategy that we call the BIG test; namely, the use of many simultaneous Bell tests in widely separated locations and using different physical systems, with each experiment's apparatus constructed and operated by different experimental teams. The only hidden-variable theories that escape this tightening are those in which choices can simultaneously influence hidden variables in many differently constructed experiments to produce a BIV in each one. This strategy is strengthened by using the same bits in many experiments, as described above.

The set of 13 BBT experiments, including true Bell tests and other realism tests requiring free choice of measurement, are summarized in Table 1 and described in Supplementary Information. Experiments (1)–(5), (8) and (11)–(13) used entangled photon pairs, (6) used entanglement between single photons and single atoms, (9) used entanglement between single photons and atoms and experiment (7) used entangled superconducting qubits. Experiments (7) and (13) used high-efficiency detection to avoid the fair-sampling assumption, thus closing simultaneously the detection-efficiency loophole and the FOCL. Experiment (5) demonstrated a violation of bilocal realism, and (10) violated a Bell inequality for multi-mode entanglement. Experiment (1) demonstrated quantum steering and (2) investigated temporal quantum correlations with a three-station measurement. Experiment (12) closed the post-selection loophole that is typically present in Bell tests based on energy–time entanglement. The analysis of the experimental results of (3) sets bounds on how well a measurement-dependent local model would have to predict Bellster behaviour to produce the observed results[30]. Experiments (3), (4), (6) and (13) tested whether human-generated measurement choices gave different results from machine-generated ones. Most experiments observed statistically strong violations of their respective inequalities, supporting the rejection of local realism in a multitude of systems and scenarios.

In summary, on 30 November 2016, a set of 13 Bell tests and similar experiments using photons, single atoms, atomic ensembles and superconducting devices, demonstrated strong disagreement with local realism, using measurement settings chosen by tens of thousands of globally distributed human participants. The results also showed empirically that measurement-setting independence—here provided by human agency—is in strong disagreement with causal determinism[17–19], a topic formerly accessible only by metaphysics. The experiments reject local realism in a wide variety of physical systems and

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Table 1 | Experiments carried out as part of the BBT, ordered by longitude, from east to west

Experiment	Lead Institution	Location	Entangled system	Rate (bps)	Inequality	Result	Stat. sig.
(1)	Griffith University	Brisbane, Australia	Photon polarization	4	$S_{16} \leq 0.511$	$S_{16} = 0.965 \pm 0.008$	57σ
(2)	University of Queensland & EQUS	Brisbane, Australia	Photon polarization	3	$\lvert S \rvert \leq 2$	$S_{AB} = 2.75 \pm 0.05$ $S_{BC} = 2.79 \pm 0.05$	15σ 16σ
(3)	USTC	Shanghai, China	Photon polarization	10^3	PRBLG[30]	$I_0 = 0.10 \pm 0.05$	N/A
(4)	IQOQI	Vienna, Austria	Photon polarization	1.61×10^3	$\lvert S \rvert \leq 2$	$S_{HRN} = 2.639 \pm 0.008$ $S_{QRN} = 2.643 \pm 0.006$	81σ 116σ
(5)	Sapienza	Rome, Italy	Photon polarization	0.62	$B \leq 1$	$B = 1.225 \pm 0.007$	32σ
(6)	LMU	Munich, Germany	Photon–atom	1.7	$\lvert S \rvert \leq 2$	$S_{HRN} = 2.427 \pm 0.0223$ $S_{QRN} = 2.413 \pm 0.0223$	19σ 18.5σ
(7)	ETHZ	Zurich, Switzerland	Transmon qubit	3×10^3	$\lvert S \rvert \leq 2$	$S = 2.3066 \pm 0.0012$	$P < 10^{-99}$
(8)	INPHYNI	Nice, France	Photon time bin	2×10^3	$\lvert S \rvert \leq 2$	$S = 2.431 \pm 0.003$	140σ
(9)	ICFO	Barcelona, Spain	Photon–atom ensemble	125	$\lvert S \rvert \leq 2$	$S = 2.29 \pm 0.10$	2.9σ
(10)	ICFO	Barcelona, Spain	Photon multi-frequency bin	20	$\lvert S \rvert \leq 2$	$S = 2.25 \pm 0.08$	3.1σ
(11)	CITEDEF	Buenos Aires, Argentina	Photon polarization	1.02	$\lvert S \rvert \leq 2$	$S = 2.55 \pm 0.07$	7.8σ
(12)	UdeC	Concepción, Chile	Photon time bin	5.2×10^4	$\lvert S \rvert \leq 2$	$S = 2.43 \pm 0.02$	20σ
(13)	NIST	Boulder, USA	Photon polarization	10^5	$K \leq 0$	$K = (1.65 \pm 0.20) \times 10^{-4}$	8.7σ

Descriptions of the experiments are given in Supplementary Information. Stat. sig., statistical significance; indicates the number of standard deviations assuming independent and identically distributed trials, unless otherwise indicated. Rate indicates the peak rate (in bits per second, bps) at which bits were used by the experiments. Owing to the limited rate of Bellster input, some experiments had dead times. B, K, S, S_{AB}, S_{BC}, S_{HRN} and S_{QRN} indicate Bell parameters for the respective experiments and S_{16} is the steering parameter (see Supplementary Information). I_0 indicates the minimum Pütz–Rosset–Barnea–Liang–Gisin measure of setting–choice independence, consistent with the observed BIV.
USTC, University of Science and Technology of China; EQUS, Centre for Engineered Quantum Systems; IQOQI, Institute for Quantum Optics and Quantum Information; INFYNI, Institut de Physique de Nice; ICFO, Institut de Ciencies Fotoniques; LMU, Ludwig-Maximilians-Universität; ETHZ, ETH Zurich; CITEDEF, Institute of Scientific and Technical Research for Defence; UdeC, University of Concepción; NIST, National Institute of Standards and Technology.

scenarios, set the groundwork for Bell-test-based applications in quantum information, introduce gamification to randomness generation and demonstrate global networking techniques by which hundreds of thousands of individuals can directly participate in experimental science.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0085-3.

Received: 31 March 2017; Accepted: 7 March 2018;
Published online 9 May 2018.

1. Bell, J. S. On the Einstein–Podolsky–Rosen paradox. *Physics* **1**, 195–200 (1964).
2. Larsson, J.-Å. Loopholes in Bell inequality tests of local realism. *J. Phys. A* **47**, 424003 (2014).
3. Kofler, J., Giustina, M., Larsson, J.-Å. & Mitchell, M. W. Requirements for a loophole-free photonic Bell test using imperfect setting generators. *Phys. Rev. A* **93**, 032115 (2016).
4. Hensen, B. et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686 (2015).
5. Giustina, M. et al. Significant-loophole-free test of Bell's theorem with entangled photons. *Phys. Rev. Lett.* **115**, 250401 (2015).
6. Shalm, L. K. et al. Strong loophole-free test of local realism. *Phys. Rev. Lett.* **115**, 250402 (2015).
7. Rosenfeld, W. et al. Event-ready Bell test using entangled atoms simultaneously closing detection and locality loopholes. *Phys. Rev. Lett.* **119**, 010402 (2017).
8. Bell, J. in *Speakable and Unspeakable in Quantum Mechanics* Ch. 7 (Cambridge Univ. Press, Cambridge UK, 2004).
9. *The BIG Bell Test* http://thebigbelltest.org.
10. Farrera, P., Heinze, G. & de Riedmatten, H. Entanglement between a photonic time-bin qubit and a collective atomic spin excitation. *Phys. Rev. Lett.* **120**, 100501 (2018).
11. Wallraff, A. et al. Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics. *Nature* **431**, 162–167 (2004).
12. Carvacho, G. et al. Experimental violation of local causality in a quantum network. *Nat. Commun.* **8**, 14775 (2017).
13. Scheidl, T. et al. Violation of local realism with freedom of choice. *Proc. Natl Acad. Sci. USA* **107**, 19708–19713 (2010).
14. Sørensen, J. et al. Exploring the quantum speed limit with computer games. *Nature* **532**, 210–213 (2016).
15. Shimony, A. *Bell's Theorem, The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) https://plato.stanford.edu/entries/bell-theorem/ (Metaphysics Research Lab, Stanford Univ., Stanford, 2017).
16. Colbeck, R. *Quantum and Relativistic Protocols for Secure Multi-Party Computation*. PhD thesis, Cambridge Univ. (2007).
17. Hoefer, C. *Causal Determinism, The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) https://plato.stanford.edu/entries/determinism-causal/ (Metaphysics Research Lab, Stanford Univ., Stanford, 2016).
18. Acín, A. & Masanes, L. Certified randomness in quantum physics. *Nature* **540**, 213–219 (2016).
19. Aaronson, S. Quantum randomness. *Am. Sci.* **102**, 266–271 (2014).
20. Abellán, C., Amaya, W., Mitrani, D., Pruneri, V. & Mitchell, M. W. Generation of fresh and pure random numbers for loophole-free Bell tests. *Phys. Rev. Lett.* **115**, 250403 (2015).
21. Fürst, M. et al. High speed optical quantum random number generation. *Opt. Express* **18**, 13029 (2010).
22. Gallicchio, J., Friedman, A. S. & Kaiser, D. I. Testing Bell's inequality with cosmic photons: closing the setting-independence loophole. *Phys. Rev. Lett.* **112**, 110405 (2014).
23. Handsteiner, J. et al. Cosmic Bell test: measurement settings from Milky Way stars. *Phys. Rev. Lett.* **118**, 060401 (2017).
24. Wu, C. et al. Random number generation with cosmic photons. *Phys. Rev. Lett.* **118**, 140402 (2017).
25. Bera, M. N., Acín, A., Mitchell, M. W. & Lewenstein, M. Randomness in quantum mechanics: philosophy, physics and technology. *Rep. Prog. Phys.* **80**, 124001 (2017).
26. Pironio, S. et al. Random numbers certified by Bell's theorem. *Nature* **464**, 1021–1024 (2010).
27. Bar-Hillel, M. & Wagenaar, W. A. The perception of randomness. *Adv. Appl. Math.* **12**, 428–454 (1991).
28. Bierhorst, P. A robust mathematical model for a loophole-free Clauser-Horne experiment. *J. Phys. A* **48**, 195302 (2015).
29. Elkouss, D. & Wehner, S. (Nearly) optimal P values for all Bell inequalities. *NPJ Quantum Inf.* **2**, 16026 (2016).
30. Pütz, G., Rosset, D., Barnea, T. J., Liang, Y.-C. & Gisin, N. Arbitrarily small amount of measurement independence is sufficient to manifest quantum nonlocality. *Phys. Rev. Lett.* **113**, 190402 (2014).

Acknowledgements We are grateful to the many people and organizations who contributed to this project, starting with the Bellsters (details at http://thebigbelltest.org). We thank the Departament d'Ensenyament de la Generalitat de Catalunya, Ministerio de Educación, Cultura y Deporte of Spain, CosmoCaixa, Fundación Bancaria "la Caixa", INTEF, Optical Society of America, European Centers for Outreach in Photonics (ECOP), Muncyt-Museo de Ciencia y Tecnología de Madrid, Investigación y Ciéncia, Big Van, Crea Ciència, Tencent news, Micius Salon, Politecnico di Milano, University of Waterloo–Institute for Quantum Computing, Universidad Autónoma de Barcelona, Universidad Complutense de Madrid, Università degli Studi di Milano, Toptica, EPS Young Minds, Real Sociedad Española de Física, Ajuntament de Barcelona, Ajuntament de Castelldefels, Università degli studi di Padova, Università degli studi del l'Insubria, CNRIFN Istituto di Fotonica e Nanotecnologia, Istituto d'Istruzione Superiore Carlo Livi, Webbile, and Kaitos Games. We are especially thankful for the recruitment efforts of the outreach departments at our institutions. We acknowledge financial support from CONICET; ANPCyT (Argentina), Australian Research Council Centre for Quantum Computation and Communication Technology (CE110001027, CE170100012); Australian Research Council and the University of Queensland (UQ) Centre for Engineered Quantum Systems (CE110001013, CE170100009). A.G.W. acknowledges the University of Queensland Vice-Chancellor's Research and Teaching Fellowship (Australia); Austrian Academy of Sciences (OEAW), the Austrian Science Fund (FWF) (SFB F40 (FoQuS); CoQuS No. W1210-N16), the Austrian Federal Ministry of Education, Science and Research (BMBWF) and the University of Vienna (project QUESS) (Austria), MEC; MCTIC (Brazil), Barcelona City Hall; Generalitat de Catalunya (SGR 874 and 2014-SGR-1295; CERCA programme) (Catalonia), PIA CONICYT (grants PFB0824 and PAI79160083); FONDECYT (grants 1140635, 1150101, 1160400, 3170596 and 11150325; Becas Chile); the Millennium Institute for Research in Optics (MIRO); Becas CONICYT (Chile), the National Fundamental Research Program (grant 2013CB336800); the

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

National Natural Science Foundation of China (grants 91121022, 61401441 and 61401443); the Chinese Academy of Science (Strategic Priority Research Program (B) XDB04010200); the 1000 Youth Fellowship programme; the National Natural Science Foundation of China (grant 11674193); the Science and Technology Commission of Shanghai Municipality (grant 16JC1400402) (China); ERC (grant agreements AQUMET 280169, 3DQUEST 307783, OSYRIS 339106, ERIDIAN 713682, QITBOX, QUOLAPS, QuLIMA and SuperQuNet 339871); ESA (contract number 4000112591/14/NL/US); FEDER; H2020 (QUIC 641122) and the Marie Skłodowska-Curie programme (grant agreement 748549) (European Commission); German Federal Ministry of Education and Research (projects QuOReP and Q.com-Q) (Germany); CONACyT graduate fellowship programme (Mexico), MINECO (FIS2014-60843-P, FIS2014-62181-EXP, SEV-2015-0522, FIS2015-68039-P, FIS2015-69535-R and FIS2016-79508-P; Ramon y Cajal fellowship programme; TEC2016-75080-R); the ICFOnest+ international postdoctoral fellowship programme (Spain); the Knut and Alice Wallenberg Foundation (project 'Photonic Quantum Information') (Sweden); NIST (USA), AXA Chair in Quantum Information Science; FQXi Fund; Fundació Privada CELLEX; Fundació Privada MIR-PUIG; the CELLEX-ICFO-MPQ programme; Fundació Catalunya-La Pedrera; and the International PhD-fellowship programme 'la Caixa'-Severo Ochoa.

Reviewer information *Nature* thanks S. Maniscalco (with F. Cosco, S. Hamedani Raja, H. Lyyra, J. Nokkala, B. Sokolov, N. Talarico and J. Teittinen), J. Sherson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.A., instigator; M.W.M. (contact author), project leader. Coordination, gamification and networking (ICFO): S.C., general supervision; M.M.-P., project management; M.G.-M., F.A.B. and C.A., game design; J.T., prediction engine; A.H., M.G.-M. and F.A.B., Bellster recruitment and engagement strategy, design and execution; C.A., web infrastructure and networking; M.W.M., C.A. and J.T., writing of the main manuscript with input from all authors. Experiments: (1) G.J.P., R.B.P. (contact author), F.G.J., M.M.W. and S.S., experiment design and execution. (2) A.G.W. and M.R. (contact author), experiment design and execution. (3) J.-W.P. (contact author) and Q.Z. (contact author), supervision; J.-W.P., Q.Z., X.M., X.Y. and Y.L., experiment conception and design; Z.W., L.Y., H.L. and W.Z., superconducting nanowire single-photon detector (SNSPD) fabrication and characterization; J.Z., SNSPD maintenance; M.-H.L., C.W. and Y.L., photon source design and characterization; J.-Y.G. and Y.L., software design and deployment; X.Y., protocol analysis; X.Y. and Y.L., data analysis; J.-W.P., Q.Z., C.W., X.Y. and Y.L., manuscript preparation with input from all authors. (4) T.S. (contact author), A.Z. and R.U., supervision, conception and coordination; B.L., J. Ha. and D.R., experiment execution and analysis. (5) G.C., L.S., F.A., M.B. and F.S. (contact author), experiment execution and analysis; R.C., theory support. (6) H.W., W.R. (contact author), K.R., R.G. and D.B., experiment design and execution. (7) J. He., P.K., Y.S., C.K.A., A.B., S.K., P.M., M.O., T.W., S.G., C.E. and A.W. (contact author), experiment design and execution. (8) S.T. (contact author), T.L., F.K., G.S., P.V. and O.A., experiment design and execution. (9) H.d.R. (contact author), P.F. and G.H., experiment design and execution. (10) H.d.R. (contact author), A.S., A.L., M.M., D.R., O.J. and A.M., experiment design and execution; D.C. and A. Ac., theory support. (11) M.A.L. (contact author), coordination and server communication; L.T.K., I.H.L.G. and A.G.M., experiment design and execution; C.T.S. and A.B., input data formatting; L.T.K., I.H.L.G., A.G.M., C.T.S., A.B. and M.A.L., data analysis. (12) G.X. (contact author), coordination; F.T., optical setup; P.G., A. Al., J.F., A. Cu. and G.C., optical setup support; J.C., electronics design and implementation; J.C. and F.T., experiment execution; D.M., software; G.L., P.M. and F.S., experimental support; A. Ca., theory support; J.C., E.S.G. and J.-Å.L., data analysis. (13) L.K.S. (contact author), S.N., M.S., O.S.M.-L., T.G., S.G., P.B., E.K. and R.M., experiment design, execution and analysis.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0085-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0085-3.

Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The BIG Bell Test Collaboration
C. Abellán[1], A. Acín[1,2], A. Alarcón[3,4], O. Alibart[5], C. K. Andersen[6], F. Andreoli[7], A. Beckert[6], F. A. Beduini[1], A. Bendersky[8], M. Bentivegna[7], P. Bierhorst[9], D. Burchardt[10], A. Cabello[11], J. Cariñe[3,4,12], S. Carrasco[1], G. Carvacho[7], D. Cavalcanti[1], R. Chaves[13], J. Cortés-Vega[3,12], A. Cuevas[7], A. Delgado[3,12], H. de Riedmatten[1,2], C. Eichler[6], P. Farrera[1], J. Fuenzalida[3,12,14], M. García-Matos[1], R. Garthoff[10], S. Gasparinetti[6], T. Gerrits[9], F. Ghafari Jouneghani[16,17], S. Glancy[9], E. S. Gómez[3,12], P. González[3,12], J.-Y. Guan[18,19], J. Handsteiner[14,15], J. Heinsoo[6], G. Heinze[1], A. Hirschmann[1], O. Jiménez[1], F. Kaiser[5], E. Knill[9], L. T. Knoll[20,21], S. Krinner[6], P. Kurpiers[6], M. A. Larotonda[20,21], J.-Å. Larsson[22], A. Lenhard[1], H. Li[23,24], M.-H. Li[18,19], G. Lima[3,12], B. Liu[14,25], Y. Liu[18,19], I. H. López Grande[20,21], T. Lunghi[5], X. Ma[26], O. S. Magaña-Loaiza[9], P. Magnard[6], A. Magnoni[21], M. Martí-Prieto[1], D. Martínez[3,12], P. Mataloni[7], A. Mattar[1], M. Mazzera[1], R. P. Mirin[9], M. W. Mitchell[1,2]*, S. Nam[9], M. Oppliger[6], J.-W. Pan[18,19], R. B. Patel[16,17], G. J. Pryde[16,17], D. Rauch[14,15], K. Redeker[10], D. Rieländer[1], M. Ringbauer[27,28], T. Roberson[27,28], W. Rosenfeld[10], Y. Salathé[6], L. Santodonato[7], G. Sauder[5], T. Scheidl[14,15], C. T. Schmiegelow[21], F. Sciarrino[7], A. Seri[1], L. K. Shalm[9], S.-C. Shi[29], S. Slussarenko[16,17], M. J. Stevens[9], S. Tanzilli[5], F. Toledo[3,12], J. Tura[1,30], R. Ursin[14,15], P. Vergyris[5], V. B. Verma[9], T. Walter[6], A. Wallraff[6], Z. Wang[23,24], H. Weinfurter[10,30], M. M. Weston[16,17], A. G. White[27,28], C. Wu[18,19], G. B. Xavier[3,4,22], L. You[23,24], X. Yuan[26], A. Zeilinger[14,15], Q. Zhang[18,19], W. Zhang[23,24], J. Zhong[29]

[1]ICFO - Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, Barcelona, Spain. [2]ICREA - Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. [3]Millennium Institute for Research in Optics, Universidad de Concepción, Concepción, Chile. [4]Departamento de Ingeniería Eléctrica, Universidad de Concepción, Concepción, Chile. [5]CNRS UMR 7010, Institut de Physique de Nice (INPHYNI) Université Côte d'Azur, Nice, France. [6]Department of Physics, ETH Zurich, Zurich, Switzerland. [7]Dipartimento di Fisica, Sapienza Università di Roma, Rome, Italy. [8]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and Instituto de Investigación en Ciencias de la Comunicación (ICC), CONICET, Buenos Aires, Argentina. [9]National Institute of Standards and Technology, Boulder, CO, USA. [10]Ludwig-Maximilians-Universität, Munich, Germany. [11]Departamento de Física Aplicada II, Universidad de Sevilla, Seville, Spain. [12]Departamento de Física, Universidad de Concepción, Concepción, Chile. [13]International Institute of Physics, Federal University of Rio Grande do Norte, Natal, Brazil. [14]Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Vienna, Austria. [15]Vienna Center for Quantum Science & Technology (VCQ), Faculty of Physics, University of Vienna, Vienna, Austria. [16]Centre for Quantum Computation and Communication Technology, Griffith University, Brisbane, Queensland, Australia. [17]Centre for Quantum Dynamics, Griffith University, Brisbane, Queensland, Australia. [18]Shanghai Branch, National Laboratory for Physical Sciences at Microscale and Department of Modern Physics, University of Science and Technology of China, Shanghai, China. [19]Shanghai Branch, CAS Center for Excellence and Synergetic Innovation Center in Quantum Information and Quantum Physics, University of Science and Technology of China, Shanghai, China. [20]Departamento de Investigaciones en Láseres y Aplicaciones (DEILAP), CITEDEF and CONICET, Buenos Aires, Argentina. [21]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and Insituto de Física de Buenos Aires (IFIBA), CONICET, Buenos Aires, Argentina. [22]Department of Electrical Engineering, Linköping University, Linköping, Sweden. [23]State Key Laboratory of Functional Materials for Informatics, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. [24]CAS Center for Excellence in Superconducting Electronics, Shanghai, China. [25]School of Computer, NUDT, Changsha, China. [26]Center for Quantum Information, Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. [27]Centre for Engineered Quantum Systems, University of Queensland, Brisbane, Queensland, Australia. [28]School of Mathematics and Physics, University of Queensland, Brisbane, Queensland, Australia. [29]Purple Mountain Observatory and Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Nanjing, China. [30]Max-Planck-Institut für Quantenoptik, Garching, Germany. *e-mail: morgan.mitchell@icfo.eu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Local realism, Bell parameters, Bell inequalities. In their 1935 article[31], Einstein, Podolsky and Rosen employed notions of locality (actions or observations in one location do not have immediate effects in other locations) and realism (observables have values even if we do not observe them) to argue that quantum theory was incomplete and could, in principle, be supplemented with information about which outcomes actually occur in any given run of an experiment. Bell formalized these notions by defining local hidden-variable models (LHVMs), a class of non-quantum theories that are simultaneously local and realistic. We consider the simplest case, of two systems measured by two observers, Alice and Bob. We write $x(y)$ to represent Alice's (Bob's) measurement setting, $a(b)$ to represent their measurement outcomes, and λ to represent the hidden variable—something that we cannot measure, but we include in the model to explain why a and b take on particular values. The predictions of any such bipartite LHVM are given by

$$P(a,b|x,y)=\sum_{\lambda} P(a|x,\lambda)P(b|y,\lambda)P(\lambda) \quad (1)$$

where $P(\cdot\,|\,\cdot)$ indicates a conditional probability. That is, the probability of getting outcome a and b when Alice and Bob measure x and y, respectively, can be expressed in terms of the local conditional probabilities $P(a\,|\,x,\lambda)$ and $P(b\,|\,y,\lambda)$. λ is averaged over, because of our ignorance of the value of this hidden variable. If the probabilities $P(a\,|\,x,\lambda)$ and $P(b\,|\,y,\lambda)$ are restricted to 0 or 1, then we have a deterministic LHVM, in which λ, x and y fully determine the outcomes a and b. Such LHVMs are explicitly realistic in the Einstein–Podolsky–Rosen sense. Locality is also explicit in the model. For example, $P(a\,|\,x,\lambda)$ depends on neither b nor y, so that the events at Bob's station have no influence on Alice's measurement outcome a. A mathematical notion of 'freedom' is implicit in the LHVM; x and y are included as free parameters and not, for example, as functions of λ. If $P(a\,|\,x,\lambda)$ and $P(b\,|\,y,\lambda)$ are allowed to take intermediate values, we speak of a non-deterministic LHVM. Because the unknown $P(\lambda)$ can take on intermediate values, deterministic and non-deterministic LHVMs are equivalent, and from here on we do not use this distinction.

This class of models, which by construction embody the Einstein–Podolsky–Rosen assumptions, was shown by Bell to be incapable of reproducing the predictions of quantum mechanics. For example, if Alice's and Bob's local systems are spin-1/2 particles in a singlet state, then their measurements (assumed to be ideal) will agree—that is, they will show that they are both spin-up or both spin-down—with probability

$$P(\text{agree}|\phi_a,\phi_b)\equiv P(\uparrow,\uparrow|\phi_a,\phi_b)+P(\downarrow,\downarrow|\phi_a,\phi_b)=\sin^2\left(\frac{\phi_b-\phi_a}{2}\right)$$

where $\phi_b-\phi_a$ is the angle between Alice's and Bob's analysis directions. Equation (1) cannot reproduce all the features of this distribution. No choice of $P(a\,|\,x,\lambda)$, $P(b\,|\,y,\lambda)$ and $P(\lambda)$ can give a probability $P(a,b\,|\,x,y)$ that simultaneously depends on the difference $\phi_b-\phi_a$, has high visibility (ranging from 0 to 1) and is sinusoidal.

This difference is efficiently captured by Bell inequalities. A Bell parameter is a linear combination of conditional probabilities $P(a,b\,|\,x,y)$, and a Bell inequality indicates the bounds (within the class of LHVMs) of a Bell parameter. Typically, the Bell inequalities of interest are those that are not obeyed by quantum mechanics, that is, those for which quantum correlations can be strong enough to violate the Bell inequality. Bell's theorem shows that there are such inequalities and thus that quantum mechanics cannot be 'completed' by hidden variables.

Bell inequalities also enable experimental tests of local realism. A Bell test is an experiment that makes many spatially separated measurements with varied settings to obtain $P(a,b\,|\,x,y)$ estimates that appear in a Bell parameter. If the observed Bell parameter violates the inequality, one can conclude that the measured systems are not governed by any LHVM. We note that this conclusion is always statistical and typically takes the form of a hypothesis test, leading to a conclusion of the form 'assuming nature is governed by local realism, the probability to produce the observed BIV (or a stronger one) is $P(\text{observed or stronger}\,|\,\text{local realism})\leq p$'. This p value is a key indicator of statistical significance in Bell tests.

'Freedom' in Bell tests. The use of the term 'free' to describe the choices in a Bell test derives more from mathematical usage than from its usage in philosophy, although the two are clearly related. Bell[8] (see Methods, John Stewart Bell on 'free variables') states that his use of 'free will' reflects the notion of 'free variables', that is, externally given parameters in physical theories, as opposed to dynamical variables that are determined by the mathematical equations of the theory.

Any realistic LHVM is described by equation (1). The mathematical requirements for the relevant 'freedom' are made evident by a more general description, in which the local realistic model includes also x and y and thus specifies the joint probability

$$P(a,b,x,y)=\sum_{\lambda} P(a|x,\lambda)P(b|y,\lambda)P(x,y|\lambda)P(\lambda) \quad (2)$$

Using the Kolmogorov definition of conditional probability, $P(A,B)=P(A\,|\,B)P(B)$, we find that equation (2) is reduced to equation (1), provided that $P(x,y\,|\,\lambda)=P(x,y)$, that is, provided that the settings are statistically independent of the hidden variables. According to Bayes' theorem, this condition can be written as $P(\lambda\,|\,x,y)=P(\lambda)$ and $P(x,y,\lambda)=P(x,y)P(\lambda)$. This condition is known in the literature as the freedom-of-choice assumption, although it implies more than just free choices. A more accurate term might be 'measurement setting–hidden-variable independence'. We note that this condition does not require x to be independent of y, nor does it require $P(x,y)$ to be unbiased. Similar observations emerge from the more complex calculations needed to assign p values to observed data in Bell tests[3].

The above describes the sense in which the basis choices should be 'free'. The desideratum is independence from the hidden variables that describe the particle behaviours, keeping in mind that the choices and measurements could, consistent with relativistic causality, be influenced by any event in their backward light-cones. Because the setting choices and the measurements will always have overlapping backward light-cones, it is impossible to rule out the possibility of a common past influence through space-time considerations. If human choices are free, however, such influences are excluded. We also note that complete independence is not required, although the tolerance for interdependence can be low[30,32,33]. The theory that the entire experiment, including choices and outcomes, is pre-determined by initial conditions is known as superdeterminism. Superdeterminism cannot be tested[34].

A very similar concept of 'freedom' applies to the entangled systems measured in a Bell test. A BIV with free choice and under strict locality conditions implies either indeterminacy of the measurement outcomes or faster-than-light communications and thus closed time-like curves[16,18]. If Bob's measurement outcome is predictable based on information available to him before the measurement, and if it also satisfies the condition for a BIV (namely, a strong correlation with Alice's measurement outcome that depends on his measurement choice), then Bob can influence the statistics of Alice's measurement outcome and thus communicate with her, despite being space-like separated from her. Considering that Bob could in principle have information on any event that occurs in his backward light cone, this implies (assuming no closed time-like curves) that Bob's measurement outcome must be statistically independent of all prior events.

In this way, we see that 'freedom', understood as a behaviour statistically independent of prior conditions, appears twice in a Bell test; first as a requirement on the setting choices and second as a conclusion about the nature of measurement outcomes in entangled systems. These two are linked, in that the second can be demonstrated if the first is present.

Previous tests using physical randomness generators to choose measurement settings have thus demonstrated a relationship between physical processes, showing for example[4,20] that if spontaneous emission is 'free', then the outcomes of measurements on entangled electrons are also 'free'. By using humans to make the choices, we translate this to the human realm, showing that, in the words of Conway and Kochen[35], "if indeed there exist any experimenters with a modicum of free will, then elementary particles must have their own share of this valuable commodity". Here, 'experimenters' are those who choose the settings, that is, the Bellsters. See the main text for a discussion of the locality loophole when using humans.

John Stewart Bell on 'free variables'. A brief but informative source for Bell's positions on setting choices is an exchange of opinions with Clauser, Horne and Shimony in articles titled 'The theory of local beables'[8] and 'Free variables and local causality'[36]. In the first of these articles, Bell very briefly considers using humans to choose the measurement settings

> It has been assumed [in deriving Bell's theorem] that the settings of instruments are in some sense free variables - say at the whim of experimenters - or in any case not determined in the overlap of the backward light cones.

whereas in the second article, Bell defends this choice of method and compares it against 'mechanical'—that is, physical—methods of choosing the settings

> Suppose that the instruments are set at the whim, not of experimental physicists, but of mechanical random number generators. Indeed it seems less impractical to envisage experiments of this kind…

Bell proceeds to consider the strengths and weaknesses of physical random-number generators in Bell tests, offering arguments about why under 'reasonable' assumptions, physical random-number generators might be trusted. Nonetheless, he concludes

> Of course it might be that these reasonable ideas about physical randomizers are just wrong - for the purpose at hand. A theory might appear in which such conspiracies

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

> inevitably occur, and these conspiracies may then seem more digestable than the non-localities of other theories.

In summary, Bell distinguishes different levels of persuasiveness, noting that physical setting-generators, while having the required independence in many local realistic theories, cannot be expected to do so in all such theories. In contemporary terminology, what he argues is that physical-setting generators can only tighten, not close, the FOCL.

Bell also defends his use of the concept of 'free will' in a physics context, something that had been criticized by Clauser, Horne and Shimony. Bell writes

> Here I would entertain the hypothesis that experimenters have free will […] it seems to me that in this matter I am just pursuing my profession of theoretical physics.
>
> A respectable class of theories, including contemporary quantum theory as it is practiced, have 'free' 'external' variables in addition to those internal to and conditioned by the theory. These variables provide a point of leverage for 'free willed experimenters', if reference to such hypothetical metaphysical entities is permitted. I am inclined to pay particular attention to theories of this kind, which seem to me most simply related to our everyday way of looking at the world.
>
> Of course there is an infamous ambiguity here, about just what and where the free elements are. The fields of Stern-Gerlach magnets could be treated as external. Or such fields and magnets could be included in the quantum mechanical system, with external agents acting only on the external knobs and switches. Or the external agents could be located in the brain of the experimenter. In the latter case the setting of the experiment is *not* itself a free variable. It is only more or less correlated with one, depending on how accurately the experimenter effects his intention.

It is clear from the last three sentences that Bell considers human intention—that is, human free will—to be a 'free variable' in the context of the discussion. That is, he believes human intention fulfils the assumptions of Bell's theorem, as do experimental settings faithfully derived from human intention.

Use of 'freedom-of-choice loophole' and 'locality loophole' in this work. As noted above, a statistical condition used to derive Bell's theorem is $P(x, y, \lambda) = P(x, y)P(\lambda)$, where x and y are choices and λ describes the hidden variables. This statistical condition, known as the freedom of choice assumption, does not distinguish between three possible scenarios of influence: the condition could fail if the choices influence the hidden variables, if the hidden variables influence the choices or if a third factor influences both the choices and the hidden variables[2,3,13]. According to Bayes' theorem, equivalent forms are $P(x, y \mid \lambda) = P(x, y)$, which expresses the fact that knowing λ does not give information about (x, y), and $P(\lambda \mid x, y) = P(\lambda)$, which expresses the fact that knowing (x, y) does not give information about λ. The latter relationship makes clear that influence (in either direction) is incompatible with the freedom of choice assumption. The term used for this condition should not be taken literally; the condition can be false even if the choices are fully free, in the sense of being independent of all prior conditions. This occurs, for example, if the choices are made freely but then influence the hidden variable.

By long tradition, 'locality loophole' is the term given to the possibility of influence from Alice's (Bob's) choices or measurements to Bob's (Alice's) measurement outcomes. The term 'freedom-of-choice loophole' was introduced in Scheidl et al.[13] to describe the influence of hidden variables on choices. The exact definition was "the possibility that the settings are not chosen independently from the properties of the particle pair". We note that this formulation centres on the act of choosing and its independence, which (assuming relativistic causality, an element of local realism) can only be violated by influences from past events, not future events. These loophole definitions employ the concept of influence, which is directional, to explain how the non-directional relation of independence can be broken. Similarly directional definitions have recently been applied to experiments using cosmic sources[23,24].

Our use of the term in this paper follows the definition of Scheidl et al.[13] described above: FOCL refers to the possibility of choices being influenced by any combination of hidden variables or other factors within the backward light cone of the choice, whereas the possibility of choices influencing hidden variables, which necessarily occur in the forward light cone of the choice, is included in the locality loophole. Such a division, in addition to fitting the common-sense notion of free choice, avoids counting a single possible channel of influence in both FOCL and the locality loophole.

Status of the FOCL. After recent experiments simultaneously closed the locality, detection-efficiency, memory, timing and other loopholes[4–7], the FOCL remains open. Space-time considerations can eliminate the possibility of such influence from the particles[5,6,13,37], or from other space-time regions[4,7], to the choices, but not the possibility of a sufficiently early prior influence on both the choices and the particles. To motivate freedom of choice in this scenario, well characterized physical randomizers[20,21] have been used to choose settings.

In some experiments[4–6], the physical assumption is that at least one of (i) spontaneous emission, (ii) thermal fluctuations and (iii) classical chaos[20] is uninfluenced by prior events and thus unpredictable even within local realistic theories. In other experiments[7,13,37,38], the physical assumption is that photodetection is similarly uninfluenced. While still requiring a physical assumption, and thus not closing the FOCL, this strategy tightens the loophole in various ways. First, by using space-like separation to rule out influence from certain events (for example, entangled pair creation) and from defined space-time regions. Second, by using well characterized randomness sources, for which the setting choice is known to faithfully derive from a given physical process, it avoids assumptions about the predictability of side-channel processes. Third, in the case of refs. [4–6,20], by using a physical variable that can be randomized by each of several processes, the required assumption is reduced from 'x is uninfluenced' to 'at least one of x, y and z is uninfluenced'.

Prediction engine. Generation of random sequences by humans has been studied in the field of psychology for decades[27,39]. Early studies showed that humans perform poorly when asked to produce a random sequence, choosing in a biased manner and deviating from a uniform distribution. In ref. [40] it was shown that humans playing competitive, zero-sum games that reward uniform random choices tend to produce sequences with fewer identifiable biases. One such game is matching pennies: players must simultaneously choose between heads or tails; one player wins if the results are equal and the other player wins if the results are different. This is a standard two-person game used in game theory[41] (see also ref. [42]) with a mixed-strategy Nash equilibrium: as both players try to outguess each other, by behaving randomly, they do not incentivize the other player to change their strategy.

The BIG Bell Quest reproduces the coin-matching game, with an MLA playing the part of the opponent. The MLA operates on simple principles that human players could employ; it maintains a model of the tendencies of the opponent, noting, for example, that 'after choosing 0 and 0, the opponent usually chooses 1 as the next bit'. The MLA strategy operates with very little memory, mirroring the limited short-term memory of humans.

Formally, we write $x_i \in \{0, 1\}$ for the ith input bit, $S_k \equiv \{x_1, \ldots, x_k\}$ for a sequence of k input bits, and $\boldsymbol{x}_j^{(L)} \equiv \{x_j, x_{j+1}, \ldots, x_{j+L-1}\}$ for a length-L sub-sequence of S_k starting from bit j. Given S_k as input, the algorithm predicts the value of $x_{k+1} \in \{0, 1\}$ that maximizes

$$\max_{L \in [1, L_{\max}]} f_L(x_{k+1} | \boldsymbol{x}_{k-L+1}^{(L)}) \tag{3}$$

where f estimates the probability of x following $\boldsymbol{x}$ in S_k

$$f_L(x|\boldsymbol{x}) = \frac{\#\{\boldsymbol{x}_i^{(L+1)} : \boldsymbol{x}_i^{(L+1)} = \boldsymbol{x}||x, \quad 1 \le i \le k-L\}}{\#\{\boldsymbol{x}_i^{(L)} : \boldsymbol{x}_i^{(L)} = \boldsymbol{x}, \quad 1 \le i \le k-L\}} \tag{4}$$

where || indicates concatenation and #A indicates the number of elements in set A. Equations (3) and (4) mean that the prediction algorithm identifies the most frequently input sequence, of length $L_{\max}+1$ or shorter, that the player can form when adding the bit x_{k+1}, and predicts that the player will produce the bit needed to complete that sequence. $L_{\max}$ is chosen to be 3, reflecting the limited memory of a human opponent in the coin-matching game.

In equation (4), the estimator $f_L(x \mid \boldsymbol{x})$ of the probability that x follows $\boldsymbol{x}$ in S_k is based on modelling the user's input as a Markov process[43]. The MLA keeps a running estimate $T_L(\boldsymbol{z} \mid \boldsymbol{y})$, updated with each new input bit, of the matrix describing the probabilities of transitions among length-L words, from word $\boldsymbol{y}$ to word $\boldsymbol{z}$. The estimates are simply the observed frequency of transitions in S_k. The MLA then obtains f_L as a marginal probability distribution: the probability of the first bit of $\boldsymbol{z}$ being x, conditioned on the tail of S_k being $\boldsymbol{y}$ (See Extended Data Fig. 1).

Networking strategy and architecture. The BBT required reliable, robust and scalable operation of two linked networking tasks: providing the BIG Bell Quest video game experience, as well as live aggregation and streaming of user input to the running experiments. From a networking perspective, the latter task resembles an instant messaging service, with the important asymmetry that messages from a large pool of senders (the Bellsters) are directed to a much smaller pool of recipients (the laboratories). The network architecture, shown in Fig. 1c, was implemented using Amazon Web Services IaaS (Infrastructure as a Service) products.

In the messaging component, we employed a two-layered architecture, shown in Fig. 1c. In the first layer, BBT nodes received input bits from the users and performed a real-time health check (described below) to block spamming by robot participants. The data were then sent to the second layer, a single-instance hub node that concatenated all the bits from the first stage and distributed them to the laboratories. The communication between the two layers was implemented using a memcache computation node to maximize speed and to simplify the synchronization between the two layers.

The gaming task was handled by a single layer of game nodes and a database. To protect the critical messaging task from possible attacks on the gaming components, we used separate instances to handle back-end gaming tasks, such as user

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

information and rankings, and to handle back-end tasks in the messaging chain, such as data logging. Load balancers, networking devices that distribute incoming traffic to a scalable pool of servers, were used in both the gaming and messaging front ends to avoid overloading. This design pattern is known as horizontal scaling, and is a common practice in scalable cloud systems.

This specific architecture was not available as a standard service from web service providers, but was readily constructed from standard component services. The architecture is not specific to the low-bitrate manual input collected for the BBT and could easily be adapted to other data that can be collected by personal devices, for example, audio or acceleration. The architecture was designed to solve a problem specific to time-limited projects with crowd-sourced input: owing to the single-day nature of the BBT, the unknown number and geography of the participants and the possibility of hackers or spammers, it was not practical to test the system under full-load conditions before the event itself. The two-layer architecture helped maintain all the critical servers isolated and independently operating and helped us to scale up the system smoothly when traffic increased. In the event, the traffic surpassed our initial estimates, and we deployed three additional BBT nodes at 09:00 UTC (when Europe was waking up) with no interruption of service. Such scaling up is expected to be critical for projects (for example, ref. [44]) that combine laboratory experimentation, which tends to be time-limited because of stability and resource considerations, with crowd-sourcing, which usually entails unknown and fluctuating demand.

BBT nodes. The first layer of computing resources received data from Bellsters—or, more precisely, from The BIG Bell Quest running in browsers on their computers and devices. A variable number of servers running the same software functionalities were placed behind a pre-warmed load balancer that was prepared to support up to 10,000 simultaneous connections. Users connected to the load balancer via a public URL end-point and sent the data from their browsers using websocket connections. This first layer of servers aggregated the data from each connection (that is, from each user) in independent buffers during a 0.5-s interval.

A simple but important 'health check' was performed to identify and block high-speed robotic participants. If a given user contributed more than 10 bits in a single 0.5-s interval, corresponding to a rate of more than 20 keypresses per second, the user account was flagged as being non-human, and all subsequent input from that user was removed from the data stream. No feedback was provided to the users if their account was flagged, to avoid leaking information about the blocking mechanism. This method could potentially ban honest users because of networking delays and other timing anomalies, but was necessary to prevent the greater risk of the data stream being flooded with robotic input.

Hub node. The hub node aggregated the data from all the BBT nodes and also handled the connection to the laboratories. In contrast to the BBT nodes, which had to service connections from an unknown and rapidly changing number of users, the hub node aggregated data from a small and relatively stable number of trusted instances. Overall, the two-layer design simplified the networking task of delivering input from a large and variable number of users to end points (the laboratories) receiving aggregated data streams at variable rates.

Laboratories connected to the hub requested to receive random bits from the Bellsters, which were distributed after aggregating four of the 0.5-s batches from the BBT nodes, that is, in intervals of 2 s. At the end of each interval, bits were sent to each running experiment. If an experiment had requested N bits, it was sent bits $\{x_0, x_1, \ldots, x_{N-1}\}$, that is, the earliest bits to arrive in that interval. Thus, the same bits were used simultaneously in many experiments. This helped to tighten the locality loophole, because an influence from the input bits on the measurements would have to operate in the same way in several independent experiments and in several locations. With the exception of experiments (12) and (13), the sent bits were used within the next 2-s interval. To run faster than the Bellster input rate, experiments (12) and (13) operated in burst mode, accumulating bits for a specific time and then rapidly using them. As with the BBT instances, these connections were established using websocket connections. When connecting to the hub node, the laboratories specified their bitrate requirement, which could be dynamically changed. The hub node then sent a stream of Bellster-generated bits at the requested rate. Archived bits from BBT participation prior to 30 November were cached locally at the laboratories, so that experiments could continue to run even if the flow of real-time bits was insufficient or interrupted. During the event, the flux of live bits was sufficient and no experiments used these pre-distributed bits.

Memcache node. The interface between the BBT nodes and the hub instance was implemented using a memcache node. While adding an extra computing resource slightly increased the complexity of the architecture, it added robustness and simplified operations. The memcache node, in contrast to the BBT and hub nodes, had no internet-facing functionality, making its operation less dependent on external conditions. For this reason, both the BBT nodes and the hub node were registered and maintained on the memcache node, allowing the restart of any of these internet-facing instances without loss of records or synchronization.

In addition, as detailed in the next section, there was an additional monitor node in charge of (i) recording all the random bits sent from the Bellsters to the laboratories and (ii) providing real-time feedback to the Bellsters. This functionality was isolated from the operations of the hub node. Again, by splitting the monitor and hub instances, a failure or attack in the public and non-critical real-time feedback functionality had no effect on the main, private and critical random-bit distribution task.

Monitor node. For analysis and auditing purposes, all of the bits passing through the first layer of servers were recorded in a database, together with metadata describing their origin (monitor computing resource in Fig. 1c). In particular, every bit was stored together with the username that created it and the origin timestamp. The random bitstreams sent to the individual laboratories were similarly recorded bit by bit, allowing a full reconstruction of the input to the experiments.

In post-event studies of the input data, we estimated the possible contribution from potentially machine-generated participations that were not blocked by the real-time blocking mechanism. We analysed participants whose contribution were substantial, more than 2 kbit in total, and looked for anomalous timing behaviours—such as an improbably short time spent between missions and improbably large number of bits introduced per mission, both of which are limited by the dynamics of human reactions when playing the game. By flagging participants that contributed such anomalous participations as suspicious and cross-referencing against the bits sent to the experiments, we found that no experiment received more than 0.1% of the bits from the 11 suspicious participants.

In addition to using the monitor computing resource to store all the information that was streamed to the laboratories in a database, we also implemented a real-time feedback mechanism to improve the Bellsters' participation experience. After accomplishing each mission, users were shown a report on the use of their input at each of the laboratories running at that moment, as illustrated in Extended Data Fig. 2d. The numbers shown were calculated as a binomial random process $B(n, p_i)$ with parameters $n = N$ and $p_i = R_i/R$, where N is the number of bits introduced by a user in his/her last mission, R_i is the number of bits sent to laboratory i, and R is the total number of bits entered in the last 0.5-s interval.

Gamification. The BBT required a large number of human-generated random bits in a short time, thus requiring many participants, rapid input and sustained participation. The gamification strategy was designed to maximize all of these factors. Extended Data Fig. 2 shows screenshots of the game.

While still adhering to common conventions of video games (such as levels, power-ups, boss battles, animations and sound effects), the intended appeal of The BIG Bell Quest is less its entertainment value than the opportunity to contribute to the BBT experiments and to test one's unpredictability against a computer opponent. The game design incorporated internationalization, connection to social networking, community-building features and a feedback system to inform users about their contribution to the experiment, all considered essential to attract the necessary tens of thousands of participants.

The game has a classic challenge-and-reward incentive structure. The challenge is to produce random bits while avoiding being predicted by the Oracle (see Extended Data Fig. 2a). This reproduces the 'penny-matching' game studied in psychology[41,42] and resembles the well known 'rock–paper–scissors' game, thus requiring little explanation. The Oracle is an MLA that predicts player behaviour based on patterns in past input (described in Methods, Prediction engine). Most player time was spent in a rapid 'speed game' (see Extended Data Fig. 2b), where the Bellster moves along a road by hitting 0s and 1s. This part of the game requires rapid bit generation (a few bits per second) to complete the level in time. Every 20 bits an indicator shows the player's 'unpredictability'—that is, the percentage of unpredicted bits entered thus far—and the final score reflects the number of unpredicted bits, with a power-up multiplier for bits entered during a particular time window.

The rewards are multiple: at the individual level, the player is given a score for each level (to encourage a high fraction of unguessed input) and a cumulative score (to encourage repeated play). At the community level, a sharing platform offers rankings and a tool to create groups, so that Bellsters can compare their performance among friends and colleagues. Players can also post their scores to social networks (Facebook, Twitter or Weibo) at the press of a button (see Extended Data Fig. 2e and f). At the scientific level, the game provides a report on which laboratories have used how many of a player's input and for what purpose (see Extended Data Fig. 2c). Finally, the player is occasionally rewarded with a short video pre-recorded at one of the laboratories, in which experimentalists explain a part of their experiment. User feedback (see Extended Data Fig. 2e) suggests that this approach succeeded in making Bellsters feel meaningfully involved in the project, with a positive effect on retention and propagation.

The speed game–Oracle structure was repeated through three levels of difficulty, or 'worlds' ('Users', 'Internet' and 'Laboratory'), illustrating the travel of the bits from the fingers of the Bellsters, through the internet, to the laboratories

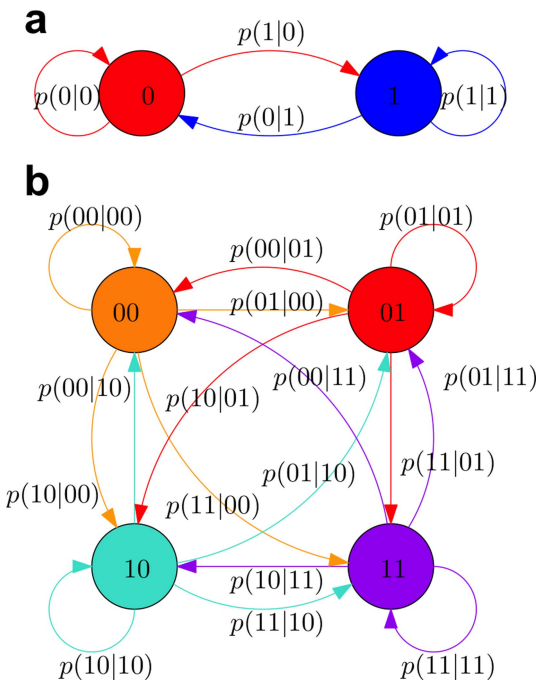
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

(see Extended Data Fig. 2c). The times, speeds and required unpredicted-input fraction in each of the levels were adjusted with the help of beta testers to avoid offering levels of trivial or impossible difficulty. The final Oracle level was objectively difficult even for an experienced player. To pass, it required $n \geq 20$ unguessed bits in a time period that allowed at most 30 bits to be entered. Even for a sequence of 30 ideal random input bits, the condition $n \geq 20$ occurs less than 5% of the time according to binomial statistics. For 30 bits that are predictable with probability of 0.6, the chance of success drops below 0.003. Nevertheless, several players persisted and completed the game (see Extended Data Fig. 2f).

Data availability. Experimental data are available upon reasonable request from the contact author of each experiment, as indicated in the author contributions. Other project data are available upon reasonable request from the corresponding author.

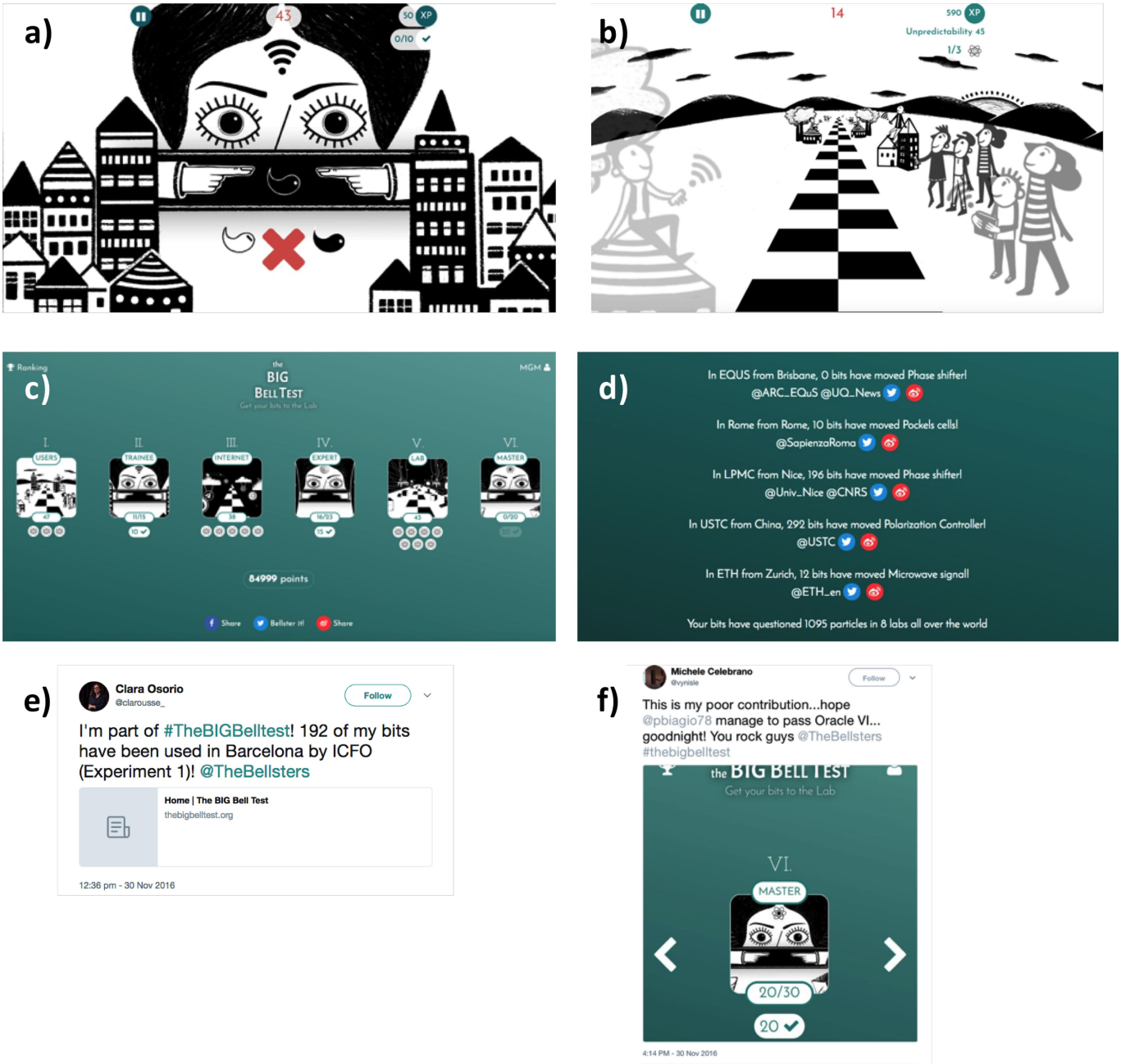
31. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47**, 777–780 (1935).
32. Hall, M. J. W. Relaxed Bell inequalities and Kochen–Specker theorems. *Phys. Rev. A* **84**, 022102 (2011).
33. Barrett, J. & Gisin, N. How much measurement independence is needed to demonstrate nonlocality? *Phys. Rev. Lett.* **106**, 100406 (2011).
34. Bell, J. S., Clauser, J., Horne, M. & Shimony, A. An exchange on local beables. *Dialectica* **39**, 85–110 (1985).
35. Conway, J. & Kochen, S. The free will theorem. *Found. Phys.* **36**, 1441–1473 (2006).
36. Bell, J. in *Speakable and Unspeakable in Quantum Mechanics* Ch. 12 (Cambridge Univ. Press, Cambridge UK, 2004).
37. Erven, C. C. et al. Experimental three-photon quantum nonlocality under strict locality conditions. *Nat. Photon.* **8**, 292–296 (2014).
38. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. & Zeilinger, A. Violation of Bell's inequality under strict Einstein locality conditions. *Phys. Rev. Lett.* **81**, 5039–5043 (1998).
39. Wagenaar, W. A. Generation of random sequences by human subjects: A critical survey of the literature. *Psychol. Bull.* **77**, 65–72 (1972).
40. Rapoport, A. & Budescu, D. V. Generation of random series in two-person strictly competitive games. *J. Exp. Psychol. Gen.* **121**, 352–363 (1992).
41. Gibbons, R. *Game theory for applied economists* (Princeton Univ. Press, Princeton, 1992).
42. Mookherjee, D. & Sopher, B. Learning behavior in an experimental matching pennies game. *Games Econ. Behav.* **7**, 62–91 (1994).
43. Serfozo, R. *Basics of applied stochastic processes* (Springer, Berlin, 2009).
44. Heck, R. et al. Remote optimization of an ultra-cold atoms experiment by experts and citizen scientists. Preprint at https://arxiv.org/abs/1709.02230 (2017).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 1 | Markov chains. a, Markov chain for $L = 1$. States are represented by circles of different colours, and transitions between states by arrows coloured as the initial state. The last input bit determines the state of the predictor. The probability $p(a \mid b)$ of a transition from state b to state a is estimated from the sequence S_k. **b**, Markov chain for $L = 2$, with four states determined by the last two input bits. The transitions model the probability $p(a, b \mid c, d)$ that the user will input bits ab, given that the last two bits were cd. The final prediction is based on the marginal of the next single bit, which is extracted from these estimated probabilities.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 2 | Screenshots from The BIG Bell Quest, illustrating various game elements. a, The Oracle uses an MLA to predict user input. **b**, The 'running' component of the game, in which participants are asked to enter a minimum number of bits with a minimum unpredicted-input fraction in a limited time. **c**, Sequence of increasing-difficulty levels, interspersed with Oracle challenges. **d**, In-game feedback on the use of the user's input bits in the running experiments. Blue and red buttons allow instant sharing on social networks Twitter and Weibo, respectively. **e**, A social media post sharing participant results. **f**, A social media post by a participant who completed the very difficult last Oracle level. The BIG Bell Quest artwork by Maria Pascual (Kaitos Games). See also Methods, Gamification.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0082-6

Highly durable, coking and sulfur tolerant, fuel-flexible protonic ceramic fuel cells

Chuancheng Duan[1]*, Robert J. Kee[1], Huayang Zhu[1], Canan Karakaya[1], Yachao Chen[1], Sandrine Ricote[1], Angelique Jarry[2], Ethan J. Crumlin[3], David Hook[4], Robert Braun[1], Neal P. Sullivan[1] & Ryan O'Hayre[1]*

Protonic ceramic fuel cells, like their higher-temperature solid-oxide fuel cell counterparts, can directly use both hydrogen and hydrocarbon fuels to produce electricity at potentially more than 50 per cent efficiency[1,2]. Most previous direct-hydrocarbon fuel cell research has focused on solid-oxide fuel cells based on oxygen-ion-conducting electrolytes, but carbon deposition (coking) and sulfur poisoning typically occur when such fuel cells are directly operated on hydrocarbon- and/or sulfur-containing fuels, resulting in severe performance degradation over time[3–6]. Despite studies suggesting good performance and anti-coking resistance in hydrocarbon-fuelled protonic ceramic fuel cells[2,7,8], there have been no systematic studies of long-term durability. Here we present results from long-term testing of protonic ceramic fuel cells using a total of 11 different fuels (hydrogen, methane, domestic natural gas (with and without hydrogen sulfide), propane, *n*-butane, *i*-butane, iso-octane, methanol, ethanol and ammonia) at temperatures between 500 and 600 degrees Celsius. Several cells have been tested for over 6,000 hours, and we demonstrate excellent performance and exceptional durability (less than 1.5 per cent degradation per 1,000 hours in most cases) across all fuels without any modifications in the cell composition or architecture. Large fluctuations in temperature are tolerated, and coking is not observed even after thousands of hours of continuous operation. Finally, sulfur, a notorious poison for both low-temperature and high-temperature fuel cells, does not seem to affect the performance of protonic ceramic fuel cells when supplied at levels consistent with commercial fuels. The fuel flexibility and long-term durability demonstrated by the protonic ceramic fuel cell devices highlight the promise of this technology and its potential for commercial application.

Without any modifications in composition or architecture of the protonic ceramic fuel cells (PCFCs), we achieve excellent performance with direct operation (that is, no fuel pre-processing, pre-reforming or conditioning) on 11 different fuel streams (Fig. 1a), many of which have not previously been studied in a PCFC. Our PCFC operating directly from the flame tip of a butane lighter (Supplementary Video 1) further highlights the fuel flexibility and thermal shock resistance of these cells. Peak power densities at 600 °C (Fig. 1a, b) range from 0.66 W cm^{-2} for H_2 to 0.17 W cm^{-2} for iso-octane (data at lower temperatures provided in Extended Data Fig. 1). Direct PCFC operation on NH_3 and CH_3OH fuels approaches the performance of pure H_2, suggesting that these fuels could be particularly attractive for commercial PCFC applications. Performance on simulated domestic natural gas (a mixture of 95 vol% CH_4, 3 vol% C_2H_6 and 1 vol% C_3H_8, balanced with N_2) exceeds the performance on pure methane and is insensitive to the presence of 19.5 p.p.m. H_2S impurity, slightly higher than the maximum total sulfur concentration of 17 p.p.m. permitted for US domestic natural gas supply. Moreover, PCFC performance is not severely degraded even after 1,000 h of continuous operation on H_2S-contaminated natural gas (Fig. 1f).

Performance on the linear hydrocarbon fuels (alkanes) follows the trend $CH_4 < C_3H_8 < C_4H_{10}$: that is, performance increases with increasing hydrocarbon chain length. In contrast, performance on alcohol fuels follows the inverse trend $CH_3OH > C_2H_5OH$. The temperature required for propane and butane steam reforming is lower than methane steam reforming because the C–H bond strength decreases with increasing hydrocarbon chain length (for example, from 435 kJ mol^{-1} for CH_4 to 412 kJ mol^{-1} for C_3H_8 to 409 kJ mol^{-1} for C_4H_{10}). In fact, reforming literature suggests that propane and butane conversion can approach 100% at 450 °C in the presence of appropriate catalysts[9]. The enhanced low-temperature reforming rates of propane and butane therefore contribute considerably to the enhanced PCFC performance on these fuels relative to methane. Gas chromatography analysis (Extended Data Fig. 2a, b) of the PCFC anode outlet stream under propane and butane operation, respectively, shows that H_2 yield from C_4H_{10} is higher than that from C_3H_8 at 600 °C, matching the PCFC performance trends. Considering the alcohol fuels, the presence of a C–C bond in ethanol (which methanol lacks) greatly hinders reforming and can lead to unfavourable, coke-prone intermediates such as ethane and ethylene[10]. These factors probably contribute to the superior PCFC performance observed for methanol relative to ethanol and are consistent with previous observations in solid-oxide fuels cells (SOFCs)[11].

In most cases, we conducted PCFC testing at inlet steam-to-carbon ratios, S:C (or oxygen-to-carbon ratios, O:C), just outside the graphitic carbon deposition region based on equilibrium thermodynamic calculations[12] (Fig. 1c). To aid comparison, we tested all fuels at similar O:C ratios, targeting O:C values of about 2.5:1 which are within the range that can be realistically met using commercially feasible SOFC balance-of-system designs[13]. For the higher hydrocarbon fuels (propane, butane and iso-octane), the target O:C ratio was difficult to achieve using steam alone; in these cases, we also supplied some O_2 to the fuel stream to meet the target O:C ratios. Previous analyses have shown that thermally sustaining operation of PCFCs with direct internal steam-methane reforming at S:C ratios of 2.5:1 is feasible[14]. Irreversible heat generation from the cell is more than sufficient to offset the endotherm from in situ steam-methane reforming. In fact, direct internal steam-reforming, as conducted in this work, has additional (and substantial) process intensification benefits because the close physical co-location of the endothermic steam reforming reactions and exothermic electrochemical processes aid in rapid and efficient heat transfer.

Figure 1c illustrates a key difference between SOFC and PCFC anode mass balance when operating on hydrocarbon fuels[15,16]. During PCFC operation, hydrogen is removed from the anode via protons transiting from anode to cathode. In contrast, during SOFC operation, oxygen is introduced into the anode by oxygen ions transiting from the cathode to the anode. Thus, as the fuel is used, PCFC and SOFC anode mass balances follow different trajectories even when starting from identical compositions of the inlet fuel stream. Even though the PCFC and SOFC starting compositions are the same, the ending PCFC and SOFC anode

[1]Colorado School of Mines, Golden, CO, USA. [2]Department of Materials Science and Engineering, University of Maryland, College Park, MD, USA. [3]Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [4]CoorsTek, Golden, CO, USA. *e-mail: cduan@mines.edu; rohayre@mines.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

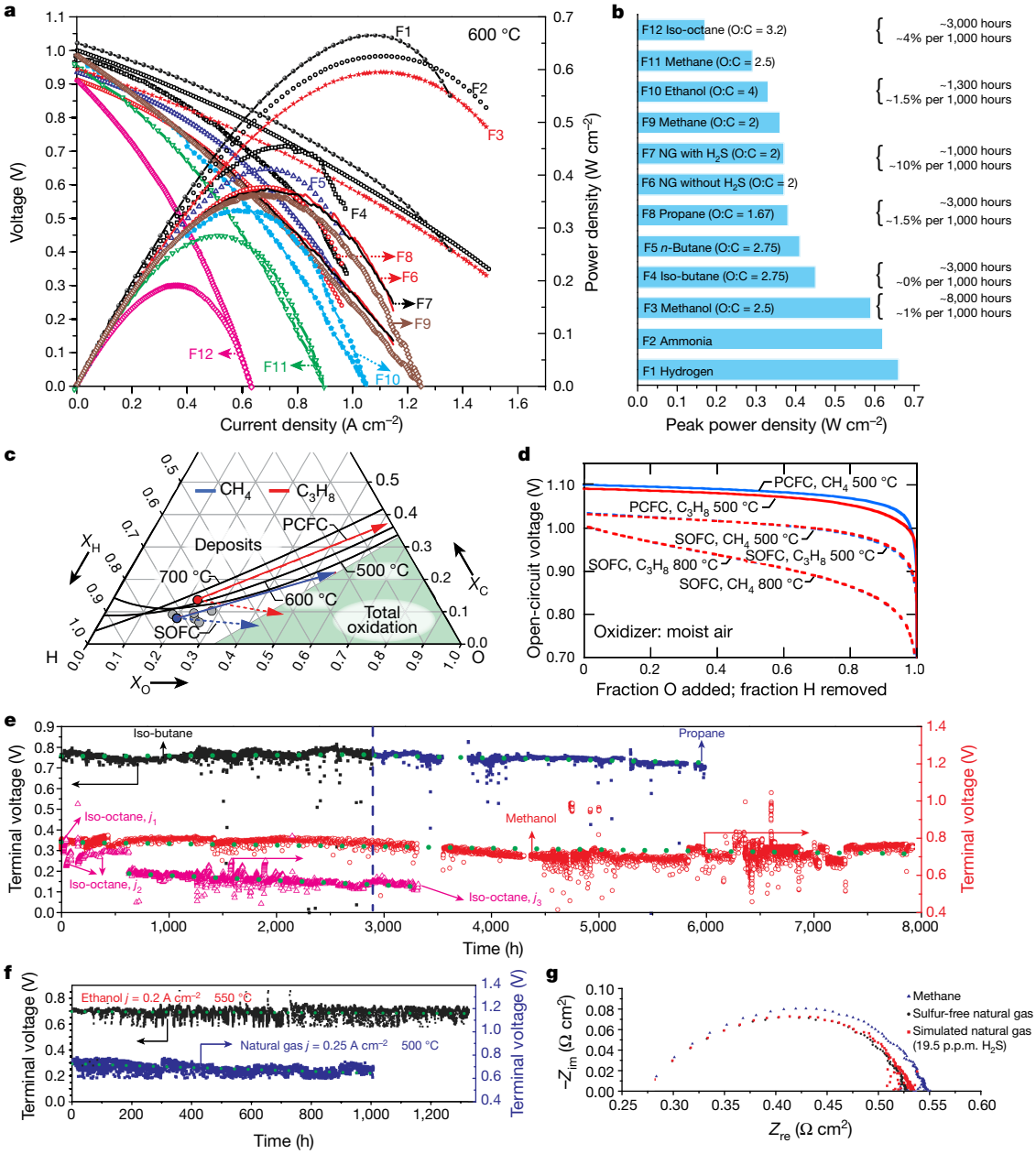


Fig. 1 | PCFC performance across a range of fuels. a, Current density versus voltage (*j*–*V*) and current density versus power density (*j*–*P*) curves of PCFCs under hydrogen (cell 1, fuel F1), ammonia (cell 2, F2), methanol (cell 3, F3), iso-butane (cell 4, F4), *n*-butane (cell 5, F5), simulated desulfurized natural gas (cell 7, F6), simulated natural gas with 19.5 p.p.m. H_2S (cell 7, F7), propane (cell 4, F8), methane (steam/carbon ratio S:C = 2, cell 7, F9), methane (S:C = 2.5, cell 8, F11), ethanol (cell 10, F10) and iso-octane (cell 9, 0.1 A cm^{-2} at 600 °C at 0.82 V, 0.2 A cm^{-2} at 600 °C at 0.73 V, 0.3 A cm^{-2} at 600 °C at 0.62 V, F12) at 600 °C. **b**, Peak power densities of PCFCs on 12 different fuel streams at 600 °C (methane is shown for two different O:C ratios). Lifetime and degradation rate of PCFCs on selected fuels also indicated. NG, natural gas. **c**, Ternary diagram shows regions of equilibrium carbon formation and full oxidation, with dots for experimental fuel compositions. Fuel composition trajectories for propane (red, 20 vol% C_3H_8 + 20 vol% O_2 + 60 vol% H_2O) and methane (blue, 33.3 vol% CH_4 + 66.7 vol% H_2O) are shown as solid lines for PCFC (indicating removing H) and dashed lines for SOFC (adding O). The end-point compositions correspond to complete oxidation of the fuel stream to CO_2 and H_2O. Thus, these mass-balance trajectories essentially correspond to the progress of the anode reaction (that is, the extent of fuel utilization). 'Deposits' indicates that carbon deposition (coking) is thermodynamically favourable in that region. **d**, OCV for propane (red) and methane (blue) fuel streams as functions of H removal for PCFC at 500 °C, and of O removal for SOFC at 500 and 800 °C. **e**, Long-term stability of direct *i*-butane/propane (cell 4; *i*-butane fuel stream composition is 10 vol% C_4H_{10} + 20 vol% O_2 + 70 vol% H_2O; propane fuel stream composition is 20 vol% C_3H_8 + 20 vol% O_2 + 60 vol% H_2O; current density for both fuels is 0.2 A cm^{-2}) and methanol (cell 3, fuel stream composition is 50 vol% methanol + 50 vol% H_2O, current density j = 0.4 A cm^{-2}) PCFCs at 500 °C and iso-octane PCFC (cell 9, j_1 = 0.1 A cm^{-2} at 600 °C at 0.79 V, j_2 = 0.2 A cm^{-2} at 600 °C at 0.73 V, j_3 = 0.3 A cm^{-2} at 600 °C at 0.62 V) at 600 °C (green dots are the linear fitting results of the degradation trend; overall degradation rate is calculated based on its slope). **f**, Long-term stability of direct-natural gas PCFC with 19.5 p.p.m. H_2S (cell 7) at 500 °C with an output current density of 0.25 A cm^{-2} and long-term stability of direct-ethanol PCFC (cell 11) at 550 °C with a constant current density of 0.2 A cm^{-2} (voltage fluctuation is due to the unstable fuel supply and high steam concentration). **g**, Nyquist plots measured at OCV of a PCFC (cell 7) under sulfur-free simulated natural gas, simulated natural gas with 19.5 p.p.m. H_2S, and methane at 600 °C. Z_{im} is the imaginary component and Z_{re} the real component of impedance.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

compositions are different because H_2O is produced in the SOFC anode (versus the cathode for PCFC). The dilution of the anode fuel stream in the SOFC leads to a decrease in the Nernst potential relative to the PCFC as a function of the extent of reaction (Fig. 1d). Thus, the PCFC can maintain higher operating voltage than a comparable SOFC, particularly at high fuel utilization[14].

Although the lack of anode H_2O production can potentially enhance PCFC performance relative to the SOFC, it also yields anode gas compositions that are more prone to coke formation. As shown in Fig. 1c, the compositions of PCFC anode gas reside within, or skirt just outside, the thermodynamic coke formation boundary across the full extent of reaction. In contrast, SOFC anode gas compositions quickly evolve away from the coke formation boundary with increasing extent of reaction. Despite the strong possibility for coking under PCFC operating conditions as indicated in Fig. 1c, we did not observe signs of coking during long-term PCFC operation on any of the tested fuels, and carbon deposits were not observed in post-mortem analyses of the cells (Extended Data Fig. 3a–f). The triple conducting oxide cathode and the electrolyte/cathode interfacial microstructures were likewise unaffected after long-term operation (Extended Data Fig. 4a, b). Several cells were continuously tested for more than 6,000 h without any sign of coking (Fig. 1e), including one cell operated for 2,800 h on *i*-butane followed by an additional 3,200 h on propane (Fig. 1e, primary *y*-axis) and a second cell operated continuously for about 8,000 h on methanol (Fig. 1e, secondary *y*-axis). In both cases, the long-term degradation rates were $<$1.5% per 1,000 h. Additional cells were tested on iso-octane (chosen as an analogue for petrol (gasoline), Fig. 1e secondary *y*-axis), natural gas (with 19.5 p.p.m. H_2S, Fig. 1f, secondary *y*-axis), and ethanol (Fig. 1f, primary *y*-axis) for $>$1,000 h each, again without observable signs of coking and with long-term degradation rates around 4% per 1,000 h, 10% per 1,000 h and 1.5% per 1,000 h, respectively. In addition, we compared the coking resistance of our PCFC anode, based on nickel/yttrium-doped barium zirconate (Ni/BZY) to a conventional SOFC anode based on nickel/yttria-stabilized zirconia (Ni/YSZ), by in situ high-temperature Raman spectroscopy (in situ HT-Raman). These studies reveal that the intrinsic coking resistance of the Ni/BZY-based PCFC anode is higher than that of a Ni/YSZ-based SOFC anode and provide insights into the mechanisms underlying this remarkable coke-resistance (Extended Data Fig. 5).

To aid direct comparison of PCFC performance on methane, desulfurized natural gas and natural gas with sulfur impurities (19.5 p.p.m. H_2S), we operated a single PCFC on all three fuel streams consecutively. As reported in Fig. 1a, b and Extended Data Fig. 1, PCFC performance on desulfurized and H_2S-containing natural gas slightly exceeds performance on pure methane. We attribute this difference to the small concentration of more-easily reformed higher hydrocarbon species (for example, C_2H_6 and C_3H_8) in real natural gas. Figure 1g provides electrochemical impedance spectroscopy Nyquist plots of this cell on all three fuels at open-circuit voltage (OCV), demonstrating that the ohmic resistance remains constant and the H_2S has no obvious effect on the electrode polarization resistance. This is in contrast to the typical behaviour in SOFCs, in which both the ohmic and electrode-polarization resistances generally increase in H_2S-contaminated fuel streams[17].

The long-term, coke-free, sulfur-tolerant, direct-fuelled fuel cell performance summarized in Fig. 1 is perhaps without precedent. SOFCs, which typically operate at temperatures above 700 °C, generally suffer from coking when operated directly on hydrocarbon fuels, even at O:C ratios higher than 2:1. Furthermore, standard consensus holds that coking should be exacerbated at lower temperatures, and the anode mass balance trajectories shown in Fig. 1c suggests that PCFCs should have greater propensity to coke than SOFCs under the same operating conditions. Nevertheless, the PCFCs examined here achieve long-term coke-free operation at temperatures considerably lower than their SOFC counterparts. Previous stability data from PCFCs and SOFCs operated on hydrocarbon/alcohol fuels reported in the literature (Extended Data Table 1) show that the PCFCs that we describe here deliver exceptional lifetime and durability. Multiple factors contribute to this behaviour, including several unique features of PCFCs that have been previously noted in the literature and several that have not.

First, a propitious in situ process of nanocatalyst formation occurs inside the PCFC anode during operation that probably contributes to durability and anti-coking resistance. In many cases, fuel cell performance increases slightly during the first several hundred hours of operation (for example, Fig. 1e), coinciding with the in situ formation of nanoscale nickel catalyst particles in the PCFC anode during the initial stages of operation under reducing conditions. Figure 2a–c demonstrates the time-dependent evolution of the exsolved Ni nanoparticle phase starting from a fresh anode (Fig. 2a), after operation on humidified methane for 300 h at 600 °C (Fig. 2b), and after operation on humidified methane for 1,400 h at 500 °C (Fig. 2c). This exsolution process leads to the formation of a relatively dense and uniform coating of Ni nanoparticles (about 5–100 nm in diameter) across the BZY electrolyte phase. Nanoparticle density increases slightly with time (compare Fig. 2b with Fig. 2c), and some agglomeration is seen, although the particles are still well dispersed and most particles remain well below 100 nm in size (Extended Data Fig. 6a, b). Transmission

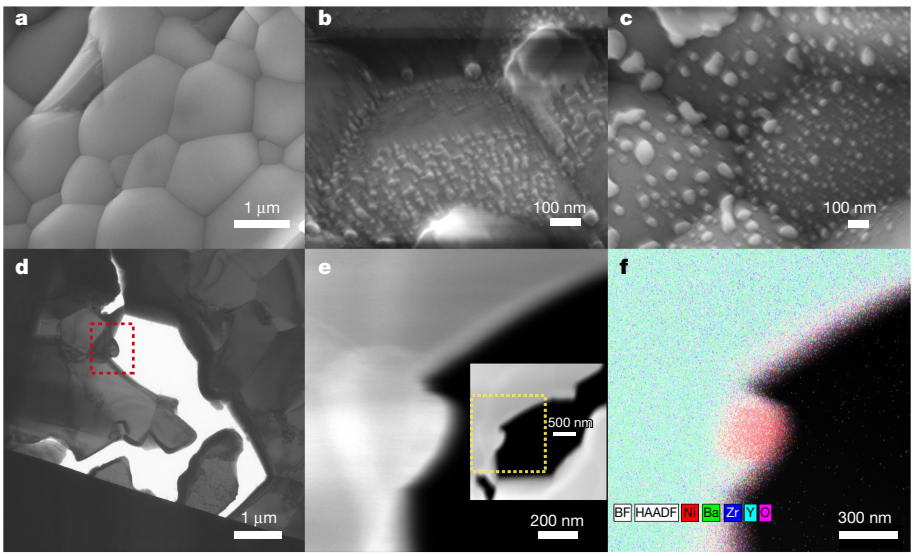


Fig. 2 | In situ exsolution of Ni nanoparticles during PCFC operation. a, Scanning electron micrograph (SEM) of BZY20 ($BaZr_{0.8}Y_{0.2}O_{3-\delta}$) PCFC anode as sintered. **b**, SEM of BZY20 PCFC anode after 300 h operation on humidified methane (28.5 vol% CH_4 + 71.5 vol% H_2O) at OCV at 600 °C. **c**, SEM of BZY20 PCFC anode after 1,400 h operation on humidified methane (20 vol% CH_4 + 50 vol% H_2O + 30 vol% Ar) under an output current density of 80 mA cm^{-2} at 500 °C. **d**, TEM micrograph of a lift-out sample from the BZY20 PCFC anode (Fig. 2c) prepared by focused ion beam (FIB). The red box is the particle on the grain boundary to be analysed in Fig. 2e, f. **e**, High-angle annular dark field (HAADF) image of a select area with one particle on the grain boundary (inset, one HAADF image with lower magnification to show the location of particle). **f**, Overlay of Ba, Zr, Y, Ni and O energy-dispersive X-ray spectroscopy (EDS) on the HAADF micrograph and bright field (BF) confirming that the nanoparticle is Ni. Panels **e**, **f** were acquired from the reverse side of the sample relative to **d**.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

electron microscopy (TEM) investigations confirm that the nanoparticles are Ni (Fig. 2d, e, f). Additional evidence in support of the in situ Ni exsolution process is provided in Extended Data Figs. 7, 8. Irvine and co-workers have recently demonstrated similar exsolution of Ni nanoparticles in $SrTiO_3$-based perovskite catalyst supports and observed greatly enhanced coking resistance for the exsolved Ni catalyst relative to comparable $SrTiO_3$ supports with deposited Ni nanoparticle catalysts[18].

As schematically illustrated in Fig. 3a, b and further discussed in the Extended Data, the unique anode microstructure resulting from

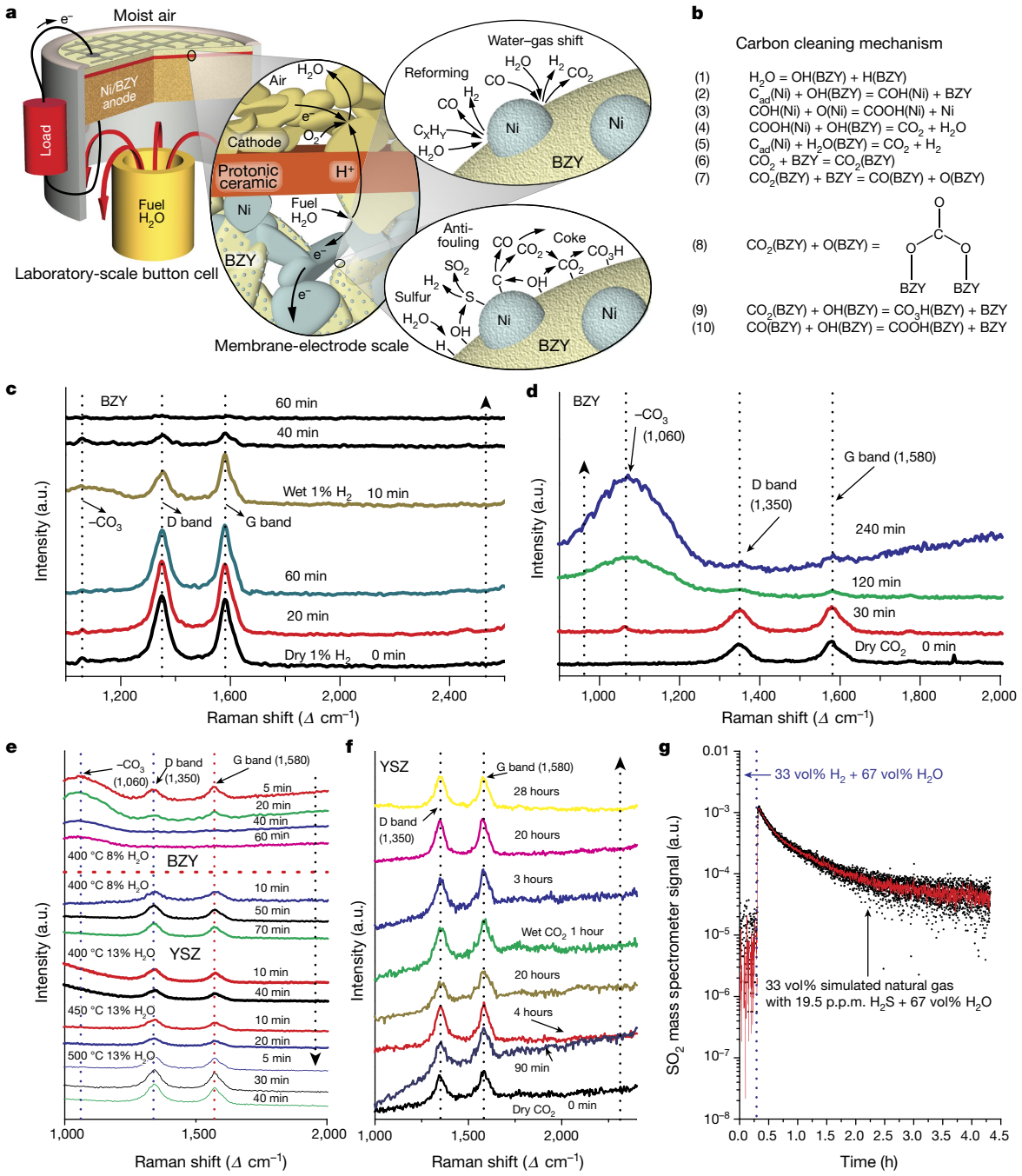


Fig. 3 | Carbon and sulfur mitigation mechanism. a, Schematic illustration of PCFC and mechanism of hydrocarbon reforming, water–gas shift reaction, sulfur cleaning and carbon cleaning. **b**, Mechanism of carbon cleaning. C_{ad} indicates carbon absorbed on the surface of Ni. **c**, In situ HT-Raman spectra of a carbon-contaminated Ni/BZY-based anode after exposure to dry 1 vol% H_2 (balance Ar) and humidified 1 vol% H_2 (10 vol% H_2O, balance Ar) at 500 °C. **d**, In situ HT-Raman spectra of a carbon-contaminated Ni/BZY-based anode after exposure to 50 vol% dry CO_2 + 49.5 vol% Ar + 0.5 vol% H_2 (dotted arrow indicates the timeline of experimental progression). **e**, In situ HT-Raman spectra of a carbon-contaminated Ni/BZY-based anode after continuous exposure to humidified 1 vol% H_2 (8 vol% H_2O) at 400 °C for 60 min; and in situ HT-Raman spectra of a carbon-contaminated Ni/YSZ-based anode after continuous exposure to humidified 1 vol% H_2 (8 vol% H_2O) at 400 °C for 70 min, then humidified 1 vol% H_2 (13 vol% H_2O) at 400 °C for 40 min, then humidified 1 vol% H_2 (13 vol% H_2O) at 450 °C for 20 min and finally humidified 1 vol% H_2 (13 vol% H_2O) at 500 °C for an additional 40 min. **f**, In situ HT-Raman spectra of a carbon-contaminated Ni/YSZ-based anode after exposure to 50 vol% dry CO_2 + 49.5 vol% Ar + 0.5 vol% H_2 for various intervals up to 20 h followed by exposure to humidified CO_2 (40 vol% CO_2 + 10 vol% H_2O + 49.5 vol% Ar + 0.5 vol% H_2) for various intervals up to an additional 28 h at 500 °C. **g**, Mass spectrometer SO_2 signal intensity as a function of time for PCFC anode during direct internal reforming of H_2S-contaminated natural gas.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

this Ni exsolution process facilitates rapid reforming chemistry with anti-fouling properties, thanks to the intimate communication between the Ni nano-islands and the supporting BZY electrolyte phase. Several researchers have observed that BZY can contribute to H_2O dissociation to form surface hydroxyl ($OH_{(s)}$) and $H_{(s)}$ compounds that can serve as surface-intermediate species in carbon removal and water–gas shift (WGS) reactions. Water dissociation can occur without a barrier on the BZY support[19]. Thus, H_2O adsorption becomes the limiting step of the $OH_{(s)}$ formation reaction, and H_2O uptake by the BZY support probably determines the carbon-removing reaction rate. At PCFC operating temperatures (500–600 °C), BZY is known to have a high H_2O-uptake capacity[20]. As $OH_{(s)}$ species are formed on the BZY, they can react with the adsorbed carbon on the Ni, suppressing coke formation[21]. Nanoscale Ni particles well dispersed on the BZY support provide high surface area and close interaction of the surface adsorbates (either between two Ni surfaces, or between Ni and BZY) by decreasing the distance for spillover (migration of adsorbates from the Ni onto the BZY and vice versa). Depending on the reaction temperature and the H_2O concentration in the feed stream, reaction between the surface carbon C(Ni) and the surface hydroxyl OH(BZY) may follow different catalytic paths and form various surface intermediates such as formyl (COH), carboxyl (COOH) or formate (HCOO)[22,23], which then further react to CO and CO_2[24]. The overall effect of this process seems to be stable catalytic behaviour even for long time-on-stream operations as observed in the Fig. 1 data.

Key in situ Raman evidence for the active role of the BZY phase in removing carbon is provided in Fig. 3c, d. A carbon-contaminated Ni/BZY anode was loaded into the in situ HT-Raman chamber under dry 1% H_2 and heated to 500 °C (Fig. 3c). Exposure to dry 1% H_2 does not decrease the intensity of the disordered and graphitic carbon Raman signals, indicating that carbon cannot be removed under dry reducing atmospheres. However, after switching to a wet reducing atmosphere, the carbon signals are eliminated within 60 min. These results strongly suggest that water plays an important role in the elimination of carbon. In addition, a Raman peak associated with -CO_3 transiently appears (and then disappears) during this carbon cleaning process in wet hydrogen, suggesting that the introduction of H_2O leads to the formation of surface -CO_3 groups that are also involved in the carbon removal process.

To further verify the above mechanism, we carried out additional in situ HT-Raman spectroscopy studies (Fig. 3d). Upon exposure to dry CO_2, we again observe the elimination of carbon in concert with the formation of -CO_3 species. In this case, the -CO_3 persists because, without H_2O, surface hydroxyls are not available for further reaction. These results demonstrate that carbon can be removed through the formation of -CO_3, whereas subsequent removal of the -CO_3 requires -OH/H_2O.

We pursued analogous experiments on a Ni/YSZ anode as summarized in Fig. 3e, f. In marked contrast to behaviour of Ni/BZY, hydration, hydroxylation and -CO_3 formation are not detected in the Ni/YSZ anode, and carbon removal does not occur upon exposure to humidified hydrogen, even when resorting to higher temperatures and/or higher levels of humidification. Furthermore, the carbon cannot be removed in dry or wet CO_2. Investigations of carbon-contaminated Ni/BZY and Ni/YSZ (Extended Data Fig. 9a, b) by in operando ambient-pressure X-ray photoelectron spectroscopy also show that Ni/BZY exhibits enhanced carbon removal compared with Ni/YSZ.

The BZY support also enhances catalyst stability under sulfur-containing hydrocarbon gas. Although sulfur is found in natural gas in various forms (thiols, COS, CS_2 and H_2S), under reforming conditions it mostly converts to H_2S, which has a well-known poisoning effect on Ni surfaces. Even H_2S concentrations as low as 1–5 p.p.m. can decrease catalytic activity by selectively adsorbing on Ni[25], and sulfur-poisoned Ni has no activity below 700 °C[26]. The sulfur uptake of the catalyst is related to the Ni surface area. Thus, Ni nanoparticles with high dispersion and high catalytic surface area can tolerate greater sulfur poisoning. Typically, sulfur can be removed through reaction with surface oxygen species to form SO_2. Over a Ni/BZY catalyst, sulfur is preferentially adsorbed on the Ni rather than the BZY support[27].

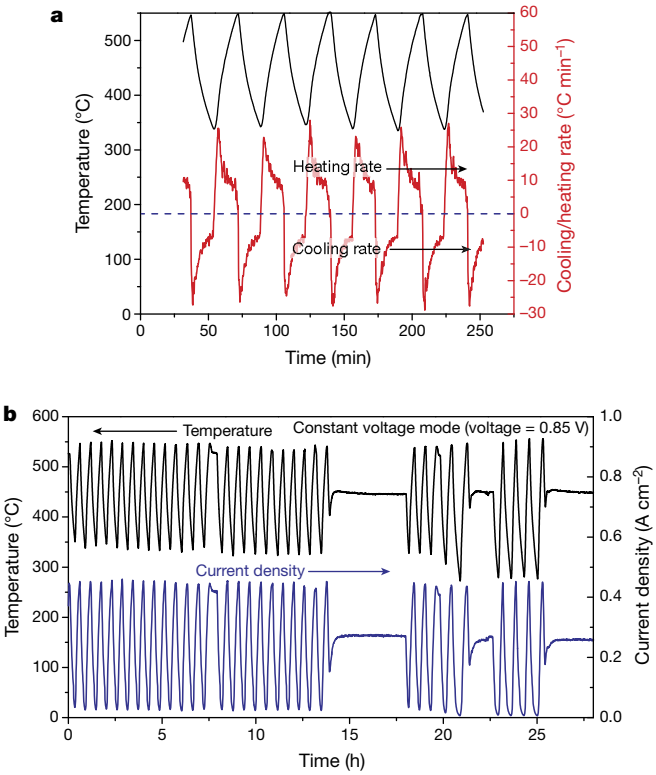


Fig. 4 | Thermal cycling measurement of hydrogen-fuelled PCFC cycling between 260 °C and 550 °C. a, Amplified temperature profile and cooling/heating rate. Heating and cooling rates reach as high as 28 °C min^{-1}. **b**, Output current was recorded at a constant voltage of 0.85 V. Based on the last cycle, the cell required 18 min to fully recover from 260 °C to its original temperature and performance at 550 °C.

Thus, the BZY support is readily available for H_2O uptake and hydroxyl formation. We speculate that this OH(BZY) can help to remove the adsorbed sulfur from the Ni surface. As indirect evidence of this reaction, we detect SO_2 production during the reforming of H_2S-containing natural gas in our PCFC anode (Fig. 3g). This is consistent with recent studies by Yang et al.[7], who observed high sulfur tolerance for $Ni/BaZr_{0.1}Ce_{0.7}Y_{0.2-x}Yb_xO_{3-\delta}$ anode materials (although their electrolyte composition was somewhat different). A series of plausible elementary reaction steps (see Methods section) can be proposed for sulfur cleaning from the Ni surface via the bi-functionality of the Ni/BZY system.

In addition to coking resistance and sulfur tolerance, thermal cycling stability is crucially important for fuel cell commercialization, particularly for applications requiring start/stop capability, or for transient or variable loads. We subjected a representative cell to 32 rapid thermal cycles while operating at a constant potential of 0.85 V (Fig. 4a, b). The temperature and current density profiles for the last thermal cycles (Fig. 4b) show that the cell power output dropped to zero as the cell temperature fell to about 260 °C. The cell required 18 min to fully recover its original temperature and performance at 550 °C. Between each thermal cycle and upon completion of the full 32 thermal cycles, the cell recovered to >99.5% of its initial current density, suggesting highly stable intrinsic thermal-cycling capability of the PCFC cell materials. The exceptional fuel flexibility and long-term durability demonstrated by the Ni/BZY-based PCFC devices reported here, combined with their thermal cycling stability, highlight the promise of this technology and its potential for commercial application.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0082-6.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Received: 12 June 2017; Accepted: 5 March 2018;
Published online 9 May 2018.

1. O'Hayre, R., Cha, S.-W., Whitney, C. & Prinz, F. B. *Fuel Cell Fundamentals* (Wiley, New York, 2009).
2. Duan, C. et al. Readily processed protonic ceramic fuel cells with high performance at low temperatures. *Science* **349**, 1321–1326 (2015).
3. Park, S., Vohs, J. M. & Gorte, R. J. Direct oxidation of hydrocarbons in a solid-oxide fuel cell. *Nature* **404**, 265–267 (2000).
4. Smith, T. R., Wood, A. & Birss, V. I. Effect of hydrogen sulfide on the direct internal reforming of methane in solid oxide fuel cells. *Appl. Catal. A* **354**, 1–7 (2009).
5. Blinn, K. S. et al. Raman spectroscopic monitoring of carbon deposition on hydrocarbon-fed solid oxide fuel cell anodes. *Energy Environ. Sci.* **5**, 7913–7917 (2012).
6. Atkinson, A. et al. Advanced anodes for high-temperature fuel cells. *Nat. Mater.* **3**, 17–27 (2004).
7. Yang, L. et al. Enhanced sulfur and coking tolerance of a mixed ion conductor for SOFCs: $BaZr_{0.1}Ce_{0.7}Y_{0.2-x}Yb_xO_{3-\delta}$. *Science* **326**, 126–129 (2009).
8. Hua, B. et al. Novel layered solid oxide fuel cells with multiple-twinned $Ni_{0.8}Co_{0.2}$ nanoparticles: the key to thermally independent CO_2 utilization and power-chemical cogeneration. *Energy Environ. Sci.* **9**, 207–215 (2016).
9. Schädel, B. T., Duisberg, M. & Deutschmann, O. Steam reforming of methane, ethane, propane, butane, and natural gas over a rhodium-based catalyst. *Catal. Today* **142**, 42–51 (2009).
10. Vaidya, P. D. & Rodrigues, A. E. Insight into steam reforming of ethanol to produce hydrogen for fuel cells. *Chem. Eng. J.* **117**, 39–49 (2006).
11. García, E. Y. & Laborde, M. A. Hydrogen production by the steam reforming of ethanol: thermodynamic analysis. *Int. J. Hydrogen Energy* **16**, 307–312 (1991).
12. Sasaki, K. & Teraoka, Y. Equilibria in fuel cell gases. *J. Electrochem. Soc.* **150**, A885–A888 (2003).
13. Borglum, B. P. & Ghezel-Ayagh, H. Development of solid oxide fuel cells at Versa Power Systems and FuelCell Energy. *ECS Trans.* **68**, 89–94 (2015).
14. Albrecht, K. J., Duan, C., O'Hayre, R. & Braun, R. J. Modeling intermediate temperature protonic ceramic fuel cells. *ECS Trans.* **68**, 3165–3175 (2015).
15. Zhu, H. & Kee, R. J. A general mathematical model for analyzing the performance of fuel-cell membrane-electrode assemblies. *J. Power Sources* **117**, 61–74 (2003).
16. Zhu, H. & Kee, R. J. Membrane polarization in mixed-conducting ceramic fuel cells and electrolyzers. *Int. J. Hydrogen Energy* **41**, 2931–2943 (2016).
17. Li, M. et al. Enhancing sulfur tolerance of Ni-based cermet anodes of solid oxide fuel cells by ytterbium-doped barium cerate infiltration. *ACS Appl. Mater. Interfaces* 8, (10293–10301 (2016).
18. Neagu, D. et al. Nano-socketed nickel particles with enhanced coking resistance grown in situ by redox exsolution. *Nat. Commun.* **6**, 8120 (2015).
19. Bandura, A. V., Evarestov, R. A. & Kuruch, D. D. Hybrid HF–DFT modeling of monolayer water adsorption on (001) surface of cubic $BaHfO_3$ and $BaZrO_3$ crystals. *Surf. Sci.* **604**, 1591–1597 (2010).
20. Liu, M. et al. Direct octane fuel cells: a promising power for transportation. *Nano Energy* **1**, 448–455 (2012).
21. Li, X. et al. In situ probing of the mechanisms of coking resistance on catalyst-modified anodes for solid oxide fuel cells. *Chem. Mater.* **27**, 822–828 (2015).
22. Karakaya, C., Otterstätter, R., Maier, L. & Deutschmann, O. Kinetics of the water–gas shift reaction over Rh/Al_2O_3 catalysts. *Appl. Catal. A* **470**, 31–44 (2014).
23. Catapan, R. C., Oliveira, A. A. M., Chen, Y. & Vlachos, D. G. DFT study of the water–gas shift reaction and coke formation on Ni(111) and Ni(211) surfaces. *J. Phys. Chem. C* **116**, 20281–20291 (2012).
24. Yang, L. et al. Promotion of water-mediated carbon removal by nanostructured barium oxide/nickel interfaces in solid oxide fuel cells. *Nat. Commun.* **2**, 357 (2011).
25. Sasaki, K. et al. H_2S poisoning of solid oxide fuel cells. *J. Electrochem. Soc.* **153**, A2023–A2029 (2006).
26. Rostrup-Nielsen, J. R., Sehested, J. & Nørskov, J. K. in *Advances in Catalysis* Vol. 47, 65–139 (2002).
27. Rostrup-Nielsen, J. & Christiansen, L. J. *Concepts in Syngas Manufacture* (Imperial College Press, London, 2011).

Acknowledgements This work was supported by Advanced Research Projects Agency–Energy (ARPA-E) for funding under the REBELS program (award DE-AR0000493). Additional support was provided by the Army Research Office under grant no. W911NF-17-1-0051 and the Colorado Office of Economic Development and International Trade (COEDIT) under their Advanced Industries Proof-of-Concept grant program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARPA-E, the Department of Energy, the Army Research Office or the US Government. This research used resources of the Advanced Light Source, which is a DOE Office of Science User Facility under contract no. DE-AC02-05CH11231. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Reviewer information *Nature* thanks J. Dailly, T. Norby, Z. Shao and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.D. and R.O'H. developed the intellectual concept, designed the fuel cell experiments, analysed the data and led the manuscript writing. N.P.S. and R.B. provided supervisory guidance on the fuel cell experiments, data interpretation, and manuscript refinement. Y.C. and C.D. conducted the in situ HT-Raman experiments. S.R., A.J. and E.J.C. designed, executed and analysed the Advanced Light Source ambient pressure X-ray photoelectron spectroscopy studies. D.H. performed the high-temperature X-ray diffraction experiments to study the phase formation of BaY_2NiO_5. C.K., H.Z. and R.J.K. developed the PCFC versus SOFC anode mass balance and Nernst potential model comparison and contributed to the discussion and analysis of the bi-functional anti-coking and sulfur-resistant mechanism hypotheses ascribed to the Ni/BZY anode.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0082-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0082-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.D. or R.O'H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

PCFC fabrication. PCFCs were fabricated by the previously developed solid-state reactive sintering (SSRS) method. The precursor powders for the anode and electrolyte were prepared by simple mixing and drying processes. For example, the anode precursor powder of 40 wt% BZY20 + 60 wt% NiO was prepared by mixing stoichiometric amounts of $BaCO_3$ (99.8% Alfa Aesar), ZrO_2 (99.7% Alfa Aesar), Y_2O_3 (99.9% Alfa Aesar), and NiO (Ni78.5 wt%, Alfa Aesar) with 20 wt% starch pore former (based on the sum of BZY20 and NiO). This raw precursor mixture was ball-milled in isopropanol with 3-mm beads of YSZ for 48 h, followed by drying at 90 °C for 24 h. Similarly, the electrolyte precursor powder of BZY20 + 1.0 wt% NiO was prepared by mixing appropriate amounts of $BaCO_3$, ZrO_2 Y_2O_3 according to the desired BZY20 stoichiometry, with the addition of 1.0 wt% NiO as a sintering aid. This raw precursor mixture was blended using the same ball-milling and drying procedures used for the anode precursor.

In this work, the previously developed triple conducting oxide, $BaCo_{0.4}Fe_{0.4}Zr_{0.1}Y_{0.1}O_{3-\delta}$ (BCFZY0.1)[2], was used for the cathode electrode. BCFZY0.1 precursor cathode powders were synthesized by the improved sol-gel method. In brief, appropriate stoichiometric amounts of $Ba(NO_3)_2$ (Alfa Aesar), $Co(NO_3)_2{\cdot}H_2O$ (Alfa Aesar), $Fe(NO_3)_3{\cdot}9H_2O$ (Alfa Aesar), $ZrO(NO_3)_2$ 35 wt% in dilute nitric acid (Sigma Aldrich), $Y(NO_3)_3{\cdot}6H_2O$ (Alfa Aesar) and EDTA (Alfa Aesar) were dissolved in 28–30% ammonium hydroxide (Alfa Aesar) under continuous heating and stirring. Citric acid was then introduced with the molar ratio of EDTA acid: citric acid: total metal ions controlled to be around 1.5:1.5:1. Subsequently, NH_3H_2O or HNO_3 was used to adjust the pH value to around 9, and the solution became immediately clear. Upon evaporating the water, a dark purple gel was obtained. The gel was then put into a drying oven at 150 °C for 24 h to get the primary powder, which was then ball-milled with 1-butanol as solvent for 48 h. The powder was dried at 90 °C for 24 h and then calcined at 600 °C for 5 h followed by ball-milling with isopropanol as solvent for 24 h and dried at 90 °C for 24 h to obtain the final powder for cathode paste preparation. We prepared the cathode paste by mixing 4 g of the respective powders with 1.5 g BCZYYb ($BaCe_{0.7}Zr_{0.1}Y_{0.1}Yb_{0.1}O_{3-\delta}$) + 1 wt% NiO powder prepared by SSRS and calcined at 1,400 °C for 18 h, 1 g dispersant (20 wt% solsperse 28000 (Lubrizol) dissolved in 3 terpineol), and 0.3 g binder (5 wt% V-006 (Heraeus) dissolved in terpineol).

Precursor pastes of the electrolyte were prepared by mixing 30 g of the respective powders with 6 g dispersant (20 wt% solsperse 28000 (Lubrizol) dissolved in 3 terpineol), and 2 g binder (5 wt% V-006 (Heraeus) dissolved in terpineol). The anode precursor powder was dry-pressed under 375 MPa for 2 min in a circular carbon-aided steel die set with a diameter of 19 mm to produce 2-mm-thick green anode pellets. A thin electrolyte precursor paste layer (about 10–20 μm after firing) was deposited on each side of the green anode pellets by screen-printing. (The electrolyte was applied to both sides of the cell to minimize asymmetric stress issues during sintering.) This 'half-cell' anode/electrolyte structure was then fired at 1,450 °C for 18 h. After firing, the extra (stress-compensating) electrolyte layer was removed by grinding. The cathode (about 10–20 μm thick) was printed on the electrolyte by brush/screen printer and the cell was fired at 900 °C for 5 h, resulting in a complete single cell (porous anode | dense electrolyte | porous cathode) with a final sintered diameter of about 15 mm and effective active area of 0.5 cm^2.

PCFC testing. These membrane-electrode assemblies were sealed onto an alumina tube with glass powder (the sealing temperature was 900 °C) and reduced under H_2 for 24 h at 600 °C. *I–V* polarization tests of the PCFC button cells were performed with 300 ml min^{-1} dry air and 15 ml min^{-1} dry hydrogen as oxidant and fuel respectively by a Gamry Reference 3000 over a range of temperatures from 350 °C to 600 °C. The electrolyte is a mixed protonic and oxygen-ion conductor[28]. We note that even the small (for example, 2–3%) partial oxygen ion conduction expected under these operating conditions is enough to provide 3–4% steam content on the anode side of the PCFC owing to electrochemical reaction of the hydrogen with the oxygen ions. Thus, even though 'dry' hydrogen is supplied to the anode, sufficient hydration of the anode chamber is sustained. For the thermal cycling stability testing, the clamshell testing furnace was turned off and manually opened to the ambient in order to subject the cell to a rapid cooling rate. Using this approach, we achieved cooling rates as high as 28 °C min^{-1}. For all other fuels, testing was performed over a range of temperatures from 500 °C to 600 °C (or 350–600 °C where noted) with 300 ml min^{-1} (STP) air as oxidant and the following fuel stream conditions:

(1) Iso-butane: 10 vol% iso-C_4H_{10} + 20 vol% O_2 + 70 vol% H_2O (flow rate of iso-C_4H_{10}: 0.9 ml min^{-1}).

(2) Natural gas: 33.3 vol% natural gas (5 ml min^{-1}) + 66.7 vol% H_2O. The simulated natural gas was provided by Praxair Technology Inc. and had 95 vol% CH_4, 3 vol% C_2H_6, 1 vol% C_3H_8 (balanced with N_2) with or without 19.5 p.p.m. H_2S.

(3) n-butane: 10 vol% n-C_4H_{10} + 20 vol% O_2 + 70 vol% H_2O (flow rate of *n*-C_4H_{10}: 0.9 ml min^{-1}).

(4) Iso-octane: UHP nitrogen at a flow rate of 5 ml min^{-1} was flowed through a bubbler containing iso-octane with constant temperature (approximately 21 °C). The mixture of iso-octane and nitrogen (about 5.5 vol% iso-octane + 94.5 vol% nitrogen) was flowed through a second bubbler containing heated water (85 °C) giving a final inlet fuel composition of 2.3 vol% *i*-C_8H_{18} + 39.7 vol% N_2 + 58 vol% H_2O (0.29 ml min^{-1} *i*-C_8H_{18}), This produced a H_2O:C_8H_{18} ratio of 25:1 which is somewhat higher than the minimum S:C ratio of 16:1 calculated based on the stoichiometric requirements for the complete steam reforming of iso-octane.

(5) Propane: 20 vol% C_3H_8 + 20 vol% O_2 + 60 vol% H_2O (flow rate of C_3H_8: 1.6 ml min^{-1}). Testing was conducted from 500 °C to 600 °C.

(6) Methane: 28.5 vol% CH_4 (5 ml min^{-1}) + 71.5 vol% H_2O (S:C = 2.5) or 33.3 vol% CH_4 (5 ml min^{-1}) + 66.7 vol% H_2O (S:C = 2). Testing was conducted from 500 °C to 600 °C.

(7) Methanol: Fuel stream introduced as an evaporated mixture of methanol and water (molar ratio of water:methanol of 1:1, flow rate of mixed liquid is 0.024 ml min^{-1}). Testing was conducted from 350 °C to 600 °C.

(8) Ethanol: Fuel stream introduced as an evaporated mixture of ethanol and water (molar ratio of water:methanol of 7:1, flow rate of mixed liquid is 0.005 ml min^{-1}).

(9) Ammonia: 10 ml min^{-1} NH_3. Testing conducted only at 600 °C.

SOFC and PCFC open-circuit potential calculations. The open-circuit potentials calculated in Fig. 1d are based on an air flow stream with a composition of 20 vol% O_2, 77 vol% N_2 and 3 vol% H_2O. The fuel flow stream takes the equilibrium composition based on the given C, H and O fractions at the investigated temperature and pressure. However, H_2 is considered to be the only electrochemically active fuel species in the fuel stream.

Evaluation of the open circuit potential for PCFCs is not straightforward. Although the yttrium-doped barium zirconates used in most PCFCs are dominantly proton conductors, they are in fact mixed ionic–electronic conductors. Among other attributes of such electrolyte membranes, the OCV (that is, Nernst potential or equilibrium reversible potential) cannot be evaluated simply in terms of gas-phase composition differences across the electrolyte membrane.

The electronic contributions to conductivity typically take the form of small polarons. The electronic current associated with polaron mobility is often referred to as a 'leakage current', which serves to reduce the OCV relative to a membrane that is a pure ion conductor. The measurable OCV can be evaluated, but in addition to the gas-phase compositions, it depends in a more complex way upon the charge-defect mobilities within the membrane[16,29,30]. The reduced OCVs reported here do account for the mixed-conduction and leakage current.

By way of contrast, considering H_2 as the fuel, the OCVs of fuel cells based on YSZ membranes and pure protonic ceramic membranes (pure PCFC) can be evaluated as

$$E_{\mathrm{rev,YSZ}} = \frac{RT}{4F}\frac{p_{\mathrm{O_2,c}}}{p_{\mathrm{O_2,a}}}, E_{\mathrm{rev,purePCFC}} = \frac{RT}{2F}\frac{p_{\mathrm{H_2,a}}}{p_{\mathrm{H_2,c}}} \quad (1)$$

In these expressions, *p* represents the gas-phase partial pressure and subscripts 'a' and 'c' refer to anode and cathode, respectively, and *F* is Faraday's constant (96,485 C mol^{-1}). Even with the same temperature and gas-phase composition across the electrolyte membrane, the OCV may be different for pure proton and oxygen-ion conducting membranes. Depending on the particular operating conditions, OCVs for the present BZY electrolyte membranes are typically a few tens of millivolts lower than would be the case for a pure proton-conducting membrane.

High-temperature X-ray diffraction. High-temperature X-ray diffraction (XRD) analyses were performed using a Bruker D8 Discover TWIN/TWIN Diffractometer with a vertical goniometer (radius is 280 mm), Cu Kα radiation, tube voltage of 40 kV and tube current of 40 mA. The hot stage is an Anton Paar HTK 2000N and utilizes a 0.5-mm Pt strip stage. Intensities were collected in the 2 theta range between 15 and 90 with a step size of 0.02 and a measuring time of 0.5 s at each step. The heating ramp rate was 5 °C min^{-1}, with a 10-min hold at each temperature before the scan was started.

Structure and morphology characterization. X-ray diffraction (XRD) analyses of all powders and pellets were performed at room temperature using a Philips X'Pert Pro MPD diffractometer (PANalytical, Almelo, Netherlands) with Cu Kα radiation, tube voltage 45 kV and tube current 40 mA. Intensities were collected in the 2 theta range between 10° and 120° with a step size of 0.008° and a measuring time of 5 s at each step. The microstructure and chemical composition of the sintered pellets were investigated by means of Field Emission Scanning Electron Microscopy (FE-SEM, JEOL JSM7000F). TEM samples were prepared by FIB (FEI Helios Nanolab 600i). TEM and energy-dispersive spectroscopy (EDS) were performed with a FEI Talos F200X TEM/STEM.

In-situ high-temperature Raman spectroscopy measurements and discussion. Anode pellets of Ni/$BaZr_{0.8}Y_{0.2}O_{3-\delta}$ (Ni/BZY) were prepared by SSRS method using the same procedure as for the complete PCFCs. Ni/YSZ-based anode power was prepared by mixing NiO (Alfa Aesar, 60 wt%) with YSZ (Fuelcellmaterials.com) and 20 wt% starch pore former (based on the sum of YSZ and NiO). Anode pellets

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

of Ni/YSZ were prepared in the same geometry and manner as the Ni/BZY based anodes. After sintering at 1,450 °C for 18 h, both Ni/BZY and Ni/YSZ-based anode pellets were polished to 0.5 mm and reduced at 600 °C for 150 h under dry hydrogen. Some reduced pellets were quenched to room temperature under hydrogen for coking resistance measurements. To study the details of the carbon cleaning process, some reduced Ni/BZY anode and Ni/YSZ anode pellets were treated for 4 h under dry 99.9% methane at 500 °C to purposely trigger carbon deposition (coking).

A Harrick environmental chamber was used for the in situ Raman tests. Temperature was controlled by the temperature controller of the chamber. Gas compositions and flow rates were controlled by mass flow controllers. Total gas flow rate for all measurements was 50 ml min^{-1}. The gas composition of dry methane for in situ Raman tests is 5% CH_4 + 95% Ar. Steam concentration was controlled by a heated bubbler and measured by a humidity sensor.

The Raman spectroscopy measurements were conducted in situ with the Harrick environmental chamber using an Alpha 300R Confocal Raman microscope (WiTec Instruments Corp., Knoxville, TN) with a doubled Nd:YAG green laser operating at 532 nm (maximum power: 39 mW). All experiments were performed at constant laser power. An Olympus SLMPLN 20 × objective (Olympus, Center Valley, PA) with a working distance of 25 mm was used for all the measurements.

In contrast to the PCFCs presented in this work, direct hydrocarbon operation in SOFCs often results in coking even when operating with S:C ratios at or above the thermodynamically predicted carbon deposition boundary. In Extended Data Fig. 5, we compare the coking resistance of our Ni/BZY-based PCFC anode with that of a conventional Ni/YSZ-based SOFC anode by in situ high-temperature Raman spectroscopy. As shown in Extended Data Fig. 5a, carbon is detected on the Ni/YSZ-based anode after just 20 min exposure to humidified methane (S:C ratio = 2) at 500 °C. In comparison, carbon is not detected on the Ni/BZY-based PCFC anode after 360 min exposure to methane under the same conditions, nor even upon long-term (>4,000 min) exposure to humidified methane at S:C = 1 (Extended Data Fig. 5c). Instead of coking, lattice protons (OH_o*) and surface carbonate groups (-CO_3) are formed on the Ni/BZY anode under these conditions, both of which we hypothesize play a role in coke and sulfur mitigation. We attribute these differences in surface chemistry to the basic nature of BZY versus the acidic nature of YSZ. We hypothesize that a series of surface reaction processes (Fig. 3a, b main manuscript) are responsible for the remarkable coking resistance observed in the Ni/BZY anode.

Because BZY is highly basic, surface hydration and hydroxylation readily occur, thereby increasing the O:C ratio on the surface. In humidified atmospheres, absorbed carbon on the Ni or BZY surface will first react with hydroxyl groups on the BZY surface, producing CO_2 (Fig. 3b, reactions (1)–(4)). The CO_2 product can subsequently absorb on the BZY surface to form -CO_3 groups, which can subsequently react with additional -OH groups in a 'cleaning loop'. Reaction steps involving the -OH reaction intermediate are kinetically fast, and thus the steady-state surface coverage of -OH remains low.

Mechanism of sulfur and carbon cleaning. As discussed in the main paper, there are important surface-chemistry differences between PCFCs and SOFCs that probably contribute to the unique coke and sulfur-resistant properties of PCFCs. PCFC electrolytes such as BZY and yttrium-doped barium cerate (BCY) are highly basic[31]. In comparison, common SOFC electrolytes such as YSZ and gadolinium-doped ceria (GDC) are considerably more acidic. Because the electrolyte phase is present as a considerable volume fraction (20–40%) in standard Ni-based cermet (ceramic–metal composite) fuel cell anode architectures for both PCFCs and SOFCs, the resulting surface chemistry differences can greatly affect anode behaviour. Specifically, several recent SOFC studies[20,31] have demonstrated that the hydration of BZY/BCY can increase the O:C ratio on surface and at two phase boundaries with Ni which can aid carbon removal and increase the coking resistance. The presence of proton conducting ceramic particles in the anode can also inhibit the adsorption of sulfur and help with sulfur removal[17].

Density functional theory studies suggest that on Ni (111) and Ni(211) surfaces, COOH(Ni) is most likely to form among the three carbon-containing hydroxyl species. Furthermore, a previous study[23] suggests that the rate-determining step for the WGS reaction is the formation of COOH(Ni). The coke removal rate is found to be directly proportional to the COOH(Ni) formation rate as well. The plausible surface reaction steps in which COOH(Ni) is formed and consumed are highlighted below. The adsorbed species on the Ni are designated as X(Ni) whereas surface adsorbates on the BZY support are designated as X(BZY).

As the reaction pathways suggest, the bi-functionality of Ni/BZY occurs via reaction of Ni and BZY surface adsorbates. Although the mechanism is not shown in detail, the interaction of the two surface adsorbates can be explained by spillover density arguments[32]. The physical distance between the two adsorbates can determine the reaction rates.

The following elementary reaction steps are proposed for sulfur cleaning from Ni via surface hydroxyl species.

(1) $H_2S + Ni = S(Ni) + H_2$
(2) $S(Ni) + OH(BZY) = SOH(Ni) + BZY$
(3) $H_2O + Ni = H_2O(Ni)$
(4) $H_2O(Ni) = OH(Ni) + H(Ni)$
(5) $OH(Ni) + Ni = O(Ni) + H(Ni)$
(6) $SOH(Ni) + O(Ni) = SO_2(Ni) + H(Ni)$

High-temperature ambient pressure X-ray photoelectron spectroscopy measurements. Pellets of BZY10 ($BaZr_{0.9}Y_{0.1}O_{3-\delta}$) and 8YSZ ($Zr_{0.92}Y_{0.08}O_{2-\delta}$) with patterned Ni electrodes were prepared for the ambient pressure X-ray photoelectron spectroscopy (AP-XPS) measurements.

The BZY10 substrates were fabricated by SSRS. $BaCO_3$, ZrO_2 and Y_2O_3 were mixed in stoichiometric proportions with 1 wt% NiO. The mixture of powders was pelletized and sintered 10 h at 1,550 °C.

8YSZ (from Tosoh) was pressed and the specimens were also sintered for 10 h at 1,550 °C. After sintering, the pellets were polished down to 0.8 mm thickness, and Ni comb-shaped electrodes were sputtered on one side of the pellets.

High-temperature AP-XPS measurements were performed at the beamline 9.3.2 at Lawrence Berkeley National Laboratory's (LBNL) Advanced Light Source (ALS). The pellets, one at a time, were set on a boron nitrate heater, with a thermocouple placed directly onto the sample surface. The samples were heated to 450 °C under ultra-high vacuum (UHV) to clean the surface and remove adventitious carbon. In the case of the 8YSZ pellet, the sample was then cooled down to 300 °C and 100 mTorr of steam was introduced in the main chamber. During the cleaning process, low-resolution survey spectra from 650 eV to −10 eV (binding energies) were collected continuously with a beam photon energy of 710 eV. The binding energies for the collected spectra were then subsequently calibrated to the Zr 3*d* photoemission peak at 181.2 eV.

Here, the UHV cleaning behaviour of carbon-contaminated Ni/BZY (Extended Data Fig. 9a) and Ni/YSZ (Extended Data Fig. 9b) patterned microelectrode structures are compared. Both Ni/BZY and Ni/YSZ samples were subjected to in situ UHV treatment at 450 °C while AP-XPS survey spectra were continuously recorded[33,34]. Within 25 min, the adventitious carbon contamination was completely removed from the Ni/BZY sample as shown by the disappearance of the C1*s* peak at 285 eV binding energy (Extended Data Fig. 9a). In contrast, the Ni/YSZ sample still showed noticeable carbon contamination even after 80 min UHV heat treatment (Extended Data Fig. 9b). To promote the cleaning process, the Ni/YSZ sample was therefore subsequently exposed to steam (p_{H2O} = 100 mTorr) at 300 °C for an additional 150 min. Even after this treatment, however, the C1s signal was not fully eliminated, thereby underscoring the difficulty in removing carbon from a Ni/YSZ SOFC anode electrode relative to a Ni/BZY PCFC anode electrode. A more extensive surface-analysis study of Ni/BZY versus Ni/YSZ under hydrocarbon fuel streams is planned.

Sulfur mitigation experiment. We loaded 4 g anode powder, after calcination at 1,450 °C for 18 h, into a packed bed reactor. The powder was reduced under pure hydrogen at 600 °C for 100 h. After reduction, the gas composition was switched to 33 vol% H_2 + 67 vol% H_2O and the SO_2 mass spectrometer signal intensity (64 amu) was recorded with a MKS Cirrus 2 mass spectrometer. After about 18 min, the gas composition was switched to 33 vol% simulated natural gas with 19.5 p.p.m. H_2S + 67 vol% H_2O and the SO_2 mass spectrometer signal intensity (64 amu) was continuously recorded. The H_2 and natural gas flow rate were held constant at 5 ml min^{-1}.

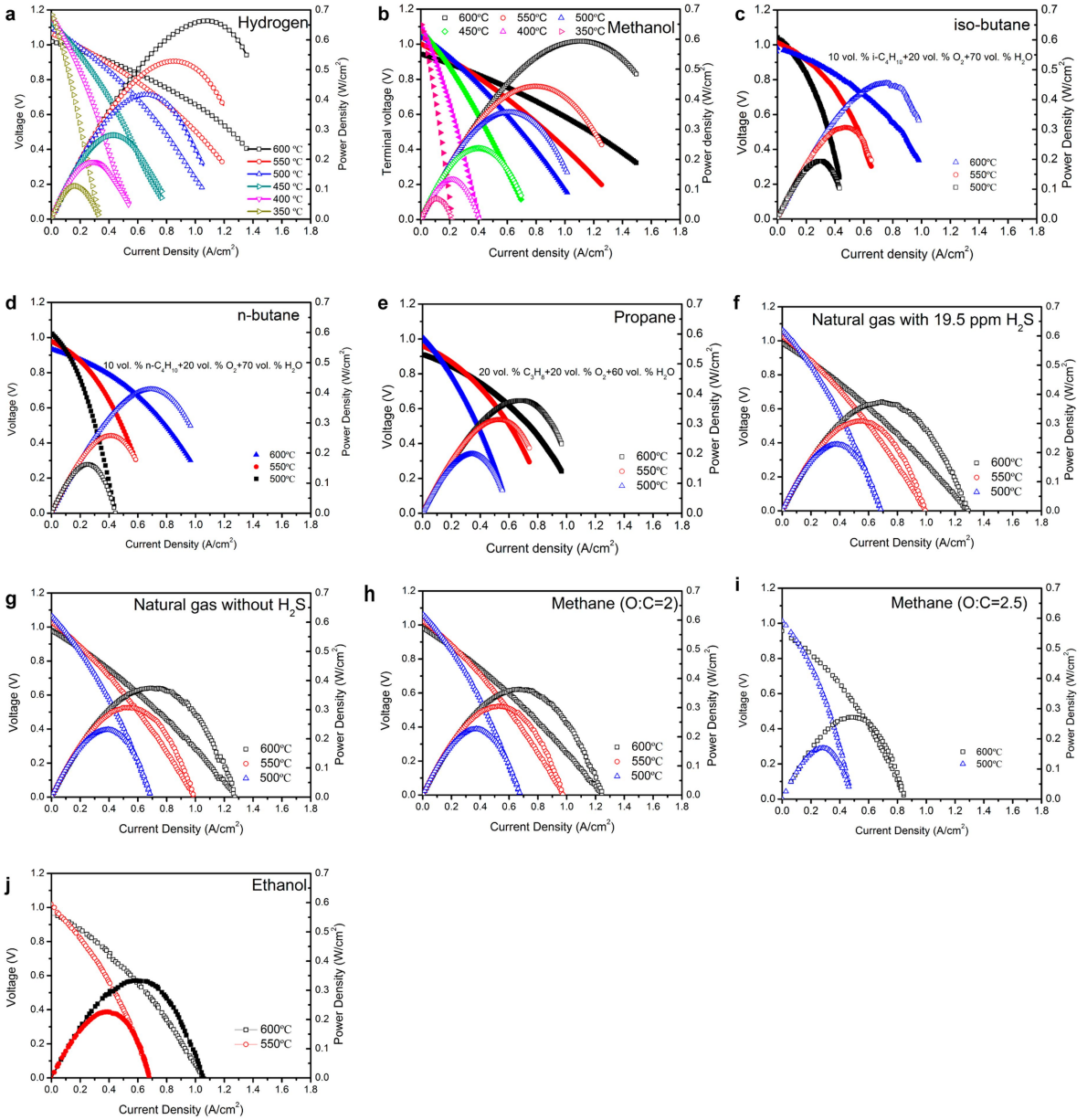
Data availability. The data that support the finding of this study are available from the authors on reasonable request. See author contributions for specific data sets.

28. Fabbri, E., Magrasó, A. & Pergolesi, D. Low-temperature solid-oxide fuel cells based on proton-conducting electrolytes. *MRS Bull.* **39**, 792–797 (2014).
29. Lai, W. & Haile, S. M. Electrochemical impedance spectroscopy of mixed conductors under a chemical potential gradient: a case study of Pt|SDC|BSCF. *Phys. Chem. Chem. Phys.* **10**, 865–883 (2008).
30. Kee, R. J. et al. Modeling the steady-state and transient response of polarized and non-polarized proton-conducting doped-perovskite membranes. *J. Electrochem. Soc.* **160**, F290–F300 (2013).
31. Shishkin, M. & Ziegler, T. Coke-tolerant Ni/$BaCe_{1-x}Y_xO_{3-\delta}$ anodes for solid oxide fuel cells: DFT + U study. *J. Phys. Chem. C* **117**, 7086–7096 (2013).
32. Pajonk, G. Contribution of spillover effects to heterogeneous catalysis. *Appl. Catal. A Gen.* **202**, 157–169 (2000).
33. Grass, M. E. et al. New ambient pressure photoemission endstation at Advanced Light Source beamline 9.3.2. *Rev. Sci. Instrum.* **81**, 053106 (2010).
34. Crumlin, E. J. et al. Surface strontium enrichment on highly active perovskites for oxygen electrocatalysis in solid oxide fuel cells. *Energy Environ. Sci.* **5**, 6081 (2012).
35. Bartholomew, C. H. Mechanisms of catalyst deactivation. *Appl. Catal. A Gen.* **212**, 17–60 (2001).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

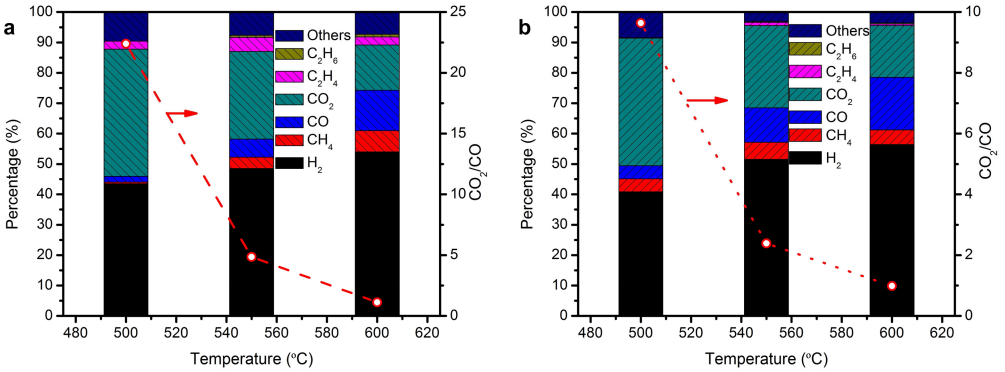
36. Tong, J. et al. Solid-state reactive sintering mechanism for large-grained yttrium-doped barium zirconate proton conducting ceramics. *J. Mater. Chem.* **20**, 6333 (2010).
37. Sengodan, S. et al. Layered oxygen-deficient double perovskite as an efficient and stable anode for direct hydrocarbon solid oxide fuel cells. *Nat. Mater.* **14**, 205–209 (2015).
38. Yang, C. et al. In situ fabrication of CoFe alloy nanoparticles structured $(Pr_{0.4}Sr_{0.6})_3(Fe_{0.85}Nb_{0.15})_2O_7$ ceramic anode for direct hydrocarbon solid oxide fuel cells. *Nano Energy* **11**, 704–710 (2015).
39. Tamm, K., Küngas, R., Gorte, R. J. & Lust, E. Solid oxide fuel cell anodes prepared by infiltration of strontium doped lanthanum vanadate into doped ceria electrolyte. *Electrochim. Acta* **106**, 398–405 (2013).
40. Yoon, D. et al. Hydrocarbon-fueled solid oxide fuel cells with surface-modified, hydroxylated Sn/Ni–$Ce_{0.8}Gd_{0.2}O_{1.9}$ heterogeneous catalyst anode. *J. Mater. Chem. A* **2**, 17041–17046 (2014).
41. Chen, Y. et al. Direct-methane solid oxide fuel cells with hierarchically porous Ni-based anode deposited with nanocatalyst layer. *Nano Energy* **10**, 1–9 (2014).
42. Tao, Z., Hou, G., Xu, N. & Zhang, Q. A highly coking-resistant solid oxide fuel cell with a nickel doped ceria: $Ce_{1-x}Ni_xO_{2-y}$ reformation layer. *Int. J. Hydrogen Energy* **39**, 5113–5120 (2014).
43. Zhan, Z. & Barnett, S. A. An octane-fueled solid oxide fuel cell. *Science* **308**, 844–847 (2005).
44. Huang, Y.-H. Double perovskites as anode materials for solid-oxide fuel cells. *Science* **312**, 254–257 (2006).
45. Choi, S. et al. A robust symmetrical electrode with layered perovskite structure for direct hydrocarbon solid oxide fuel cells: $PrBa_{0.8}Ca_{0.2}Mn_2O_{5+\delta}$. *J. Mater. Chem. A* **4**, 1747–1753 (2016).
46. Shin, T. H., Hagiwara, H., Ida, S. & Ishihara, T. RuO_2 nanoparticle-modified $(Ce,Mn,Fe)O_2$/$(La,Sr)(Fe,Mn)O_3$ composite oxide as an active anode for direct hydrocarbon type solid oxide fuel cells. *J. Power Sources* **289**, 138–145 (2015).
47. Wu, X. et al. Preparation and electrochemical performance of silver impregnated Ni-YSZ anode for solid oxide fuel cell in dry methane. *Int. J. Hydrogen Energy* **40**, 16484–16493 (2015).
48. Liu, S., Chuang, K. T. & Luo, J.-L. Double-layered perovskite anode with in situ exsolution of a Co–Fe alloy to cogenerate ethylene and electricity in a proton-conducting ethane fuel cell. *ACS Catal.* **6**, 760–768 (2016).
49. Kaur, G. & Basu, S. Physical characterization and electrochemical performance of copper-iron-ceria-YSZ anode-based SOFCs in H_2 and methane fuels. *Int. J. Energy Res.* **39**, 1345–1354 (2015).
50. Sun, Y.-F. et al. An ingenious Ni/Ce co-doped titanate based perovskite as a coking-tolerant anode material for direct hydrocarbon solid oxide fuel cells. *J. Mater. Chem. A* **3**, 22830–22838, (2015); erratum: https://doi.org/10.1039/C6TA90257J
51. Porter, D. L. & Heuer, A. H. Microstructural development in MgO-partially stabilized zirconia (Mg-PSZ). *J. Am. Ceram. Soc.* **62**, 298–305 (1979).
52. Hua, B., Li, M., Chi, B. & Jian, L. Enhanced electrochemical performance and carbon deposition resistance of Ni-YSZ anode of solid oxide fuel cells by in situ formed Ni–MnO layer for CH_4 on-cell reforming. *J. Mater. Chem. A* **2**, 1150–1158 (2014).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



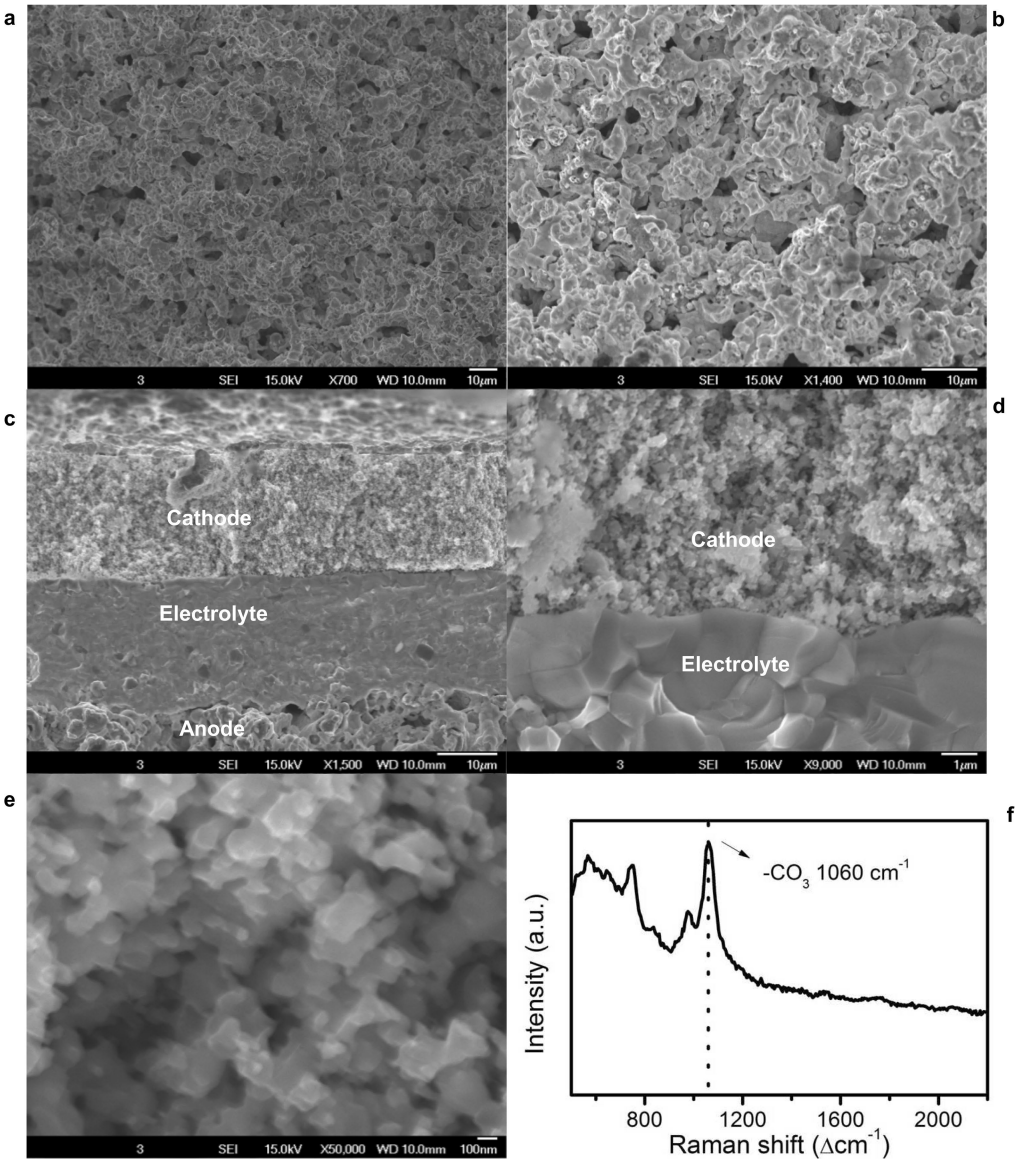
Extended Data Fig. 1 | Performance of PCFCs under 10 fuels. a, *j*–*V* and *j*–*P* curves of PCFC under hydrogen (flow rate of hydrogen: 15 ml min^{-1}, cell 1). **b**, *j*–*V* and *j*–*P* curves of direct-methanol PCFC at 600–350 °C. Flow rate of the mixture of methanol and water (molar ratio of water/ethanol is 1, cell 3) is 0.024 ml min^{-1}. **c**, *j*–*V* and *j*–*P* curves of PCFC under 10 vol% iso-C_4H_{10} + 20 vol% O_2 + 70 vol% H_2O (flow rate of iso-C_4H_{10}: 0.9 ml min^{-1}, cell 4). **d**, *j*–*V* and *j*–*P* curves of PCFC under 10 vol% *n*-C_4H_{10} + 20 vol% O_2 + 70 vol% H_2O (flow rate of *n*-C_4H_{10}: 0.9 ml min^{-1}, cell 5). **e**, *j*–*V* and *j*–*P* curves of PCFC under 20 vol% C_3H_8 + 20 vol% O_2 + 60 vol% H_2O (flow rate of C_3H_8: 1.6 ml min^{-1}, cell 4). **f**, *j*–*V* and *j*–*P* curves of PCFC under 33.3 vol% natural gas (simulated natural gas with 19.5 p.p.m. H_2S, flow rate of natural gas: 5 ml min^{-1}, cell 7) + 66.7 vol% H_2O. **g**, *j*–*V* and *j*–*P* curves of PCFC under 33.3 vol% natural gas (simulated natural gas without H_2S, flow rate of natural gas: 5 ml min^{-1}, cell 7) + 66.7 vol% H_2O. **h**, *j*–*V* and *j*–*P* curves of PCFC under 33.3 vol% methane (flow rate of methane: 5 ml min^{-1}, cell 7) + 66.7 vol% H_2O. **i**, *j*–*V* and *j*–*P* curves of PCFC under 28.5 vol% methane (flow rate of methane: 5 ml min^{-1}, cell 8) + 71.5 vol% H_2O. **j**, *j*–*V* and *j*–*P* curves of direct-ethanol PCFC at 600 °C and 550 °C. Flow rate of the mixture of ethanol and water (molar ratio of water/ethanol is 7, cell 10) is 0.005 ml min^{-1}.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



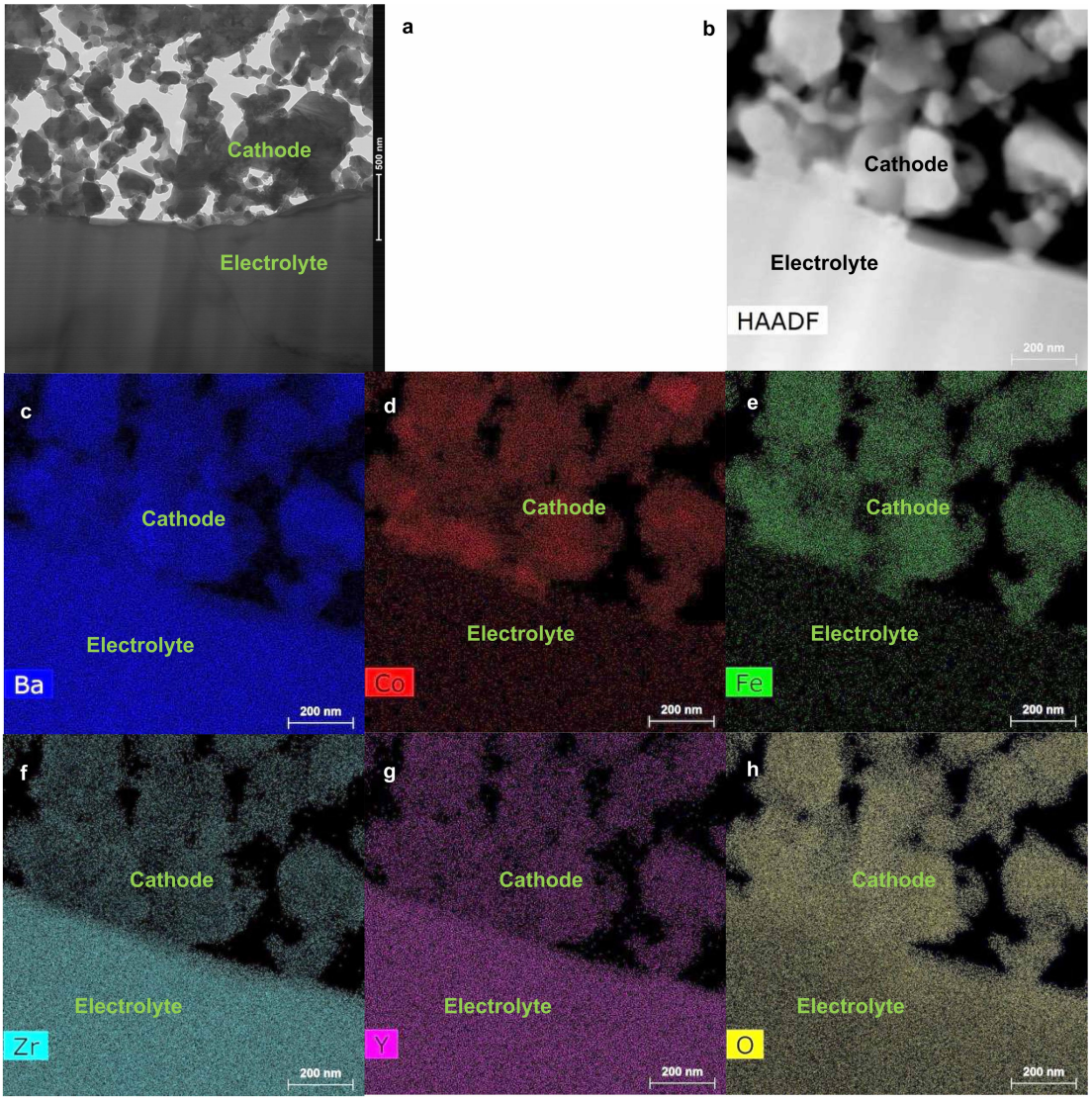
Extended Data Fig. 2 | Anode-side exhaust gas composition of PCFCs on propane and iso-butane. a, Direct-propane PCFC. **b**, Direct-iso-butane PCFC.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



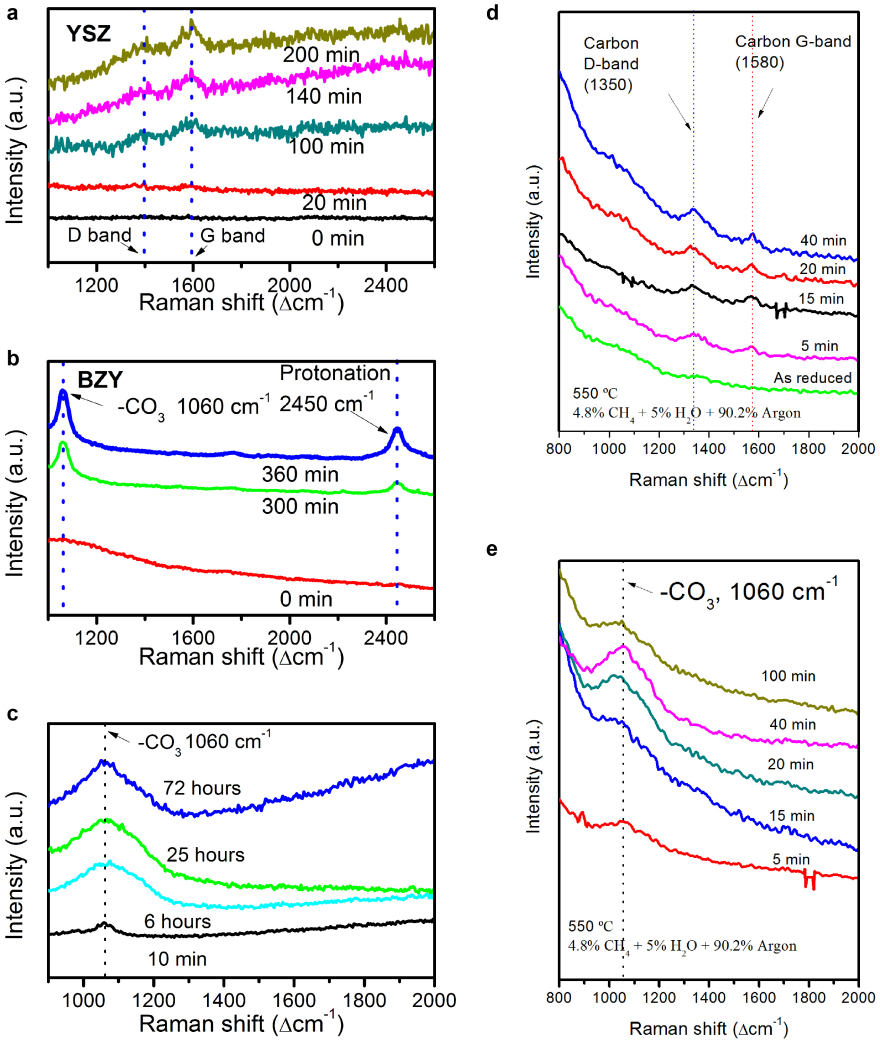
Extended Data Fig. 3 | SEM images of PCFC (Cell 4) after running for about 6,000 h on hydrocarbons at 500 °C. a, Anode (low magnification). **b**, Anode (high magnification). **c**, The sandwich structure of this cell after running for about 6,000 h. **d**, The interface between electrolyte and cathode. **e**, High magnification of cathode after 6,000 h operation. **f**, Raman spectrum of the PCFC anode after 6,000 h of operation on hydrocarbon fuel at 500 °C. A carbonate peak is visible at 1,060 cm^{-1} but the graphitic carbon (G band) and disordered carbon (D band), which would be present at 1,580 cm^{-1} and 1,350 cm^{-1}, are not apparent, indicating the absence of graphitic carbon and disordered carbon deposits in the anode even after long-term operation on hydrocarbon fuel. Long-term stabilities of direct hydrocarbons PCFCs were tested at 500 °C. At this temperature, the structure type of carbon species should be polymeric, amorphous films or filaments (C_β), vermicular filaments, fibres and/or whiskers (C_γ), or graphitic platelets or films (C_c)[35]. Typically, these carbon species are visible (by SEM) in SOFC anodes after running on hydrocarbon fuels, but the high-magnification SEM images of our PCFC anode (Extended Data Fig. 3a, b) show no visible evidence of such carbon structures. To further substantiate this conclusion, post-mortem Raman analyses show that there are no disordered and graphitic carbon species found in the PCFC anode even after long-term operation on hydrocarbon fuels (Extended Data Fig. 3f).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



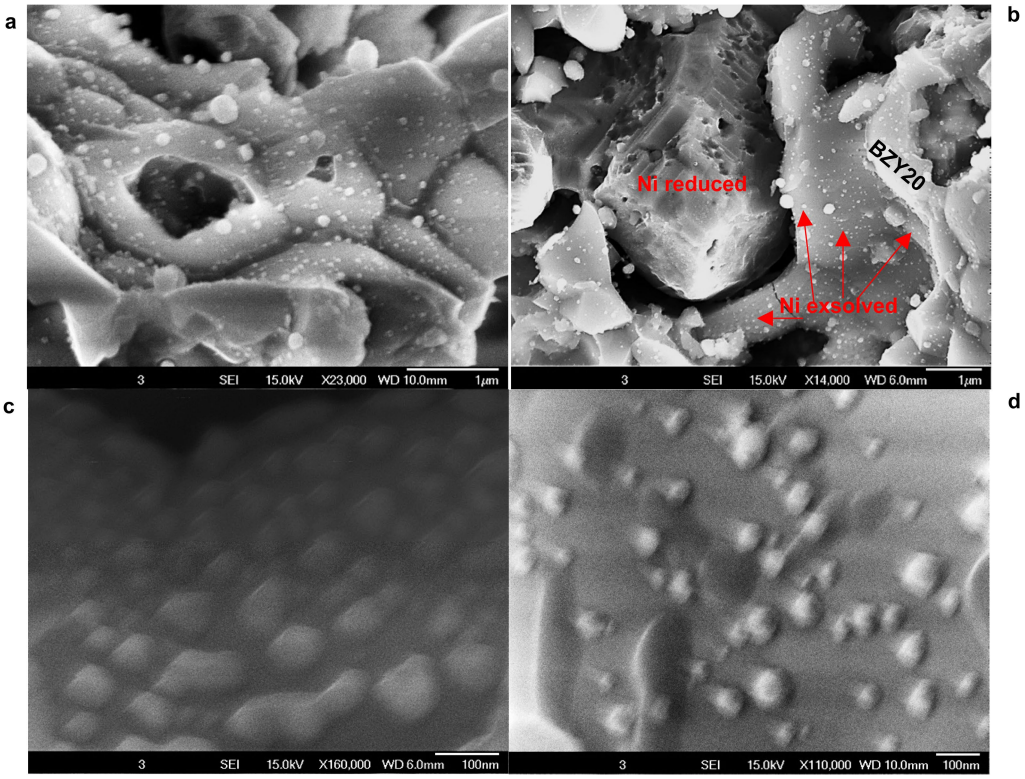
Extended Data Fig. 4 | TEM image and EDS mapping of PCFC (Cell 3) after 8,000 h operation on methanol. a, TEM image of lamella from BZY20 PCFC electrolyte/cathode interface prepared by FIB. **b**–**h**, HAADF image of the electrolyte/cathode interface, and corresponding EDS maps of Ba, Zr, Y, Co, Fe and O. There is no chemical reaction between the electrolyte and electrolyte which indicates the excellent chemical compatibility of cathode with electrolyte. EDS mapping of cathode shows there is no obvious elemental segregation.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



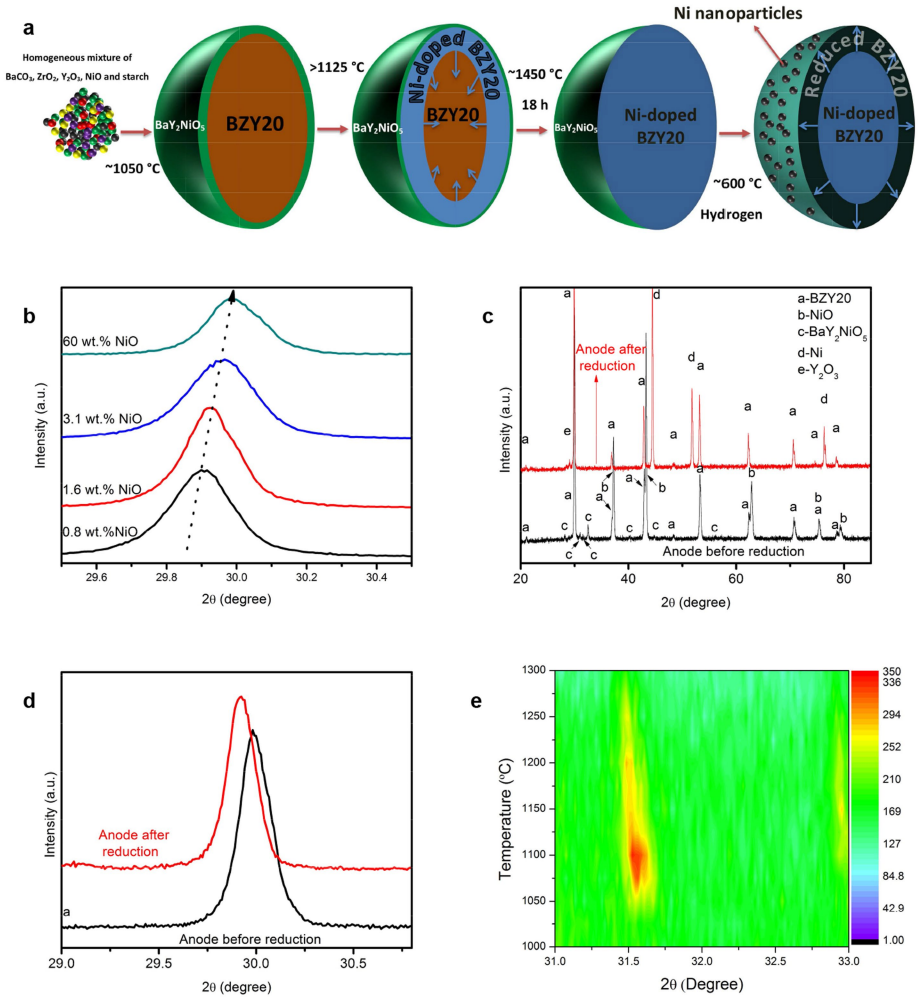
Extended Data Fig. 5 | Coking resistance of Ni/BZY and Ni/YSZ anodes investigated by in situ HT-Raman spectroscopy. a, Raman spectra of Ni/YSZ anode exposed to humidified methane (S:C = 2; 500 °C) at various times up to 200 min. **b**, Raman spectra of Ni/BZY anode exposed to humidified methane (S:C = 2; 500 °C) at various times up to 360 min. **c**, Raman spectra of Ni/BZY anode exposed to humidified methane (S:C = 1; 500 °C) at various times up to 72 h (4,320 min). **d**, Raman spectra of Ni/YSZ anode exposed to humidified methane (S:C ≈ 1; 550 °C) at various times up to 40 min. **e**, Raman spectra of Ni/BZY anode exposed to humidified methane (S:C ≈ 1; 550 °C) at various times up to 100 min.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



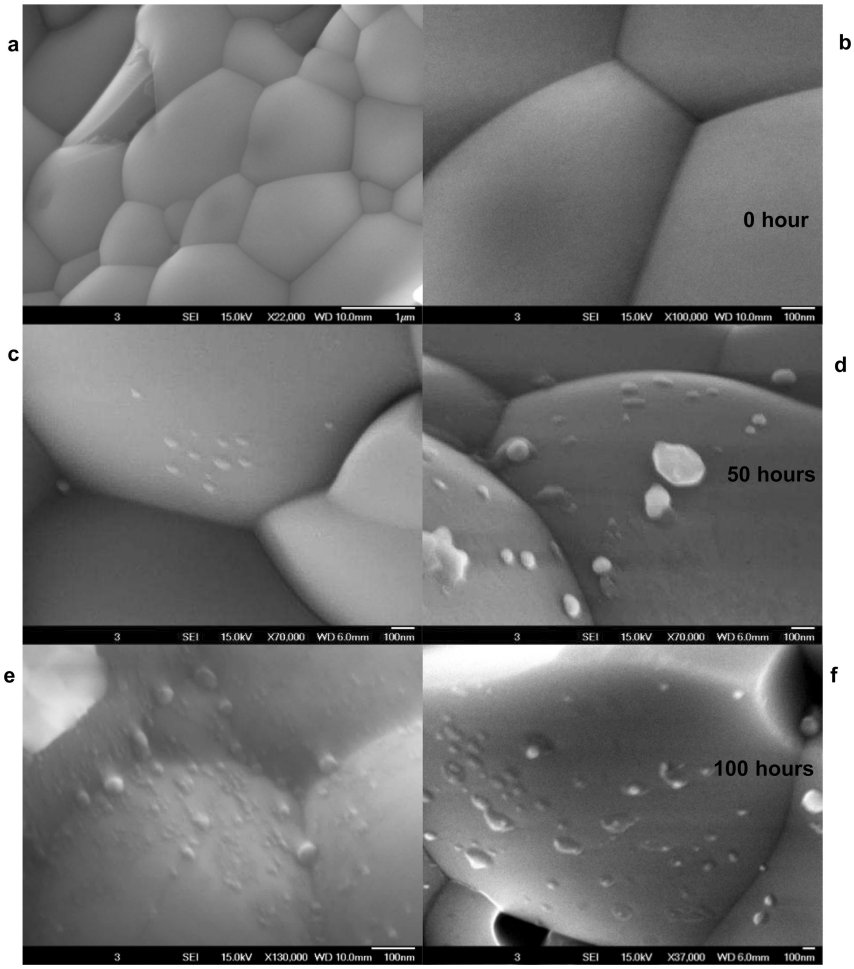
Extended Data Fig. 6 | SEM image of the anode of a PCFC after running for 1,400 and 6,000 h on hydrocarbons at 500 °C. a, **b**, Operation for 6,000 h on iso-butane and propane (cell 4). **c**, Operation for 1,400 h on methane. **d**, Operation for 6,000 h on iso-butane and propane (cell 4). The Tamman temperature of nickel is 581 °C. The target operating temperature of our PCFCs is lower than 600 °C, and most of the long-term stability testing conducted in this study was at 500 °C (that is, well below the Tamman temperature), at which Ni sintering/coarsening is avoided. Indeed, as shown in Extended Data Fig. 6c, d, the distribution and size of Ni nanoparticles on the BZY anode support remain essentially constant between 1,400 and 6,000 h of operation.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



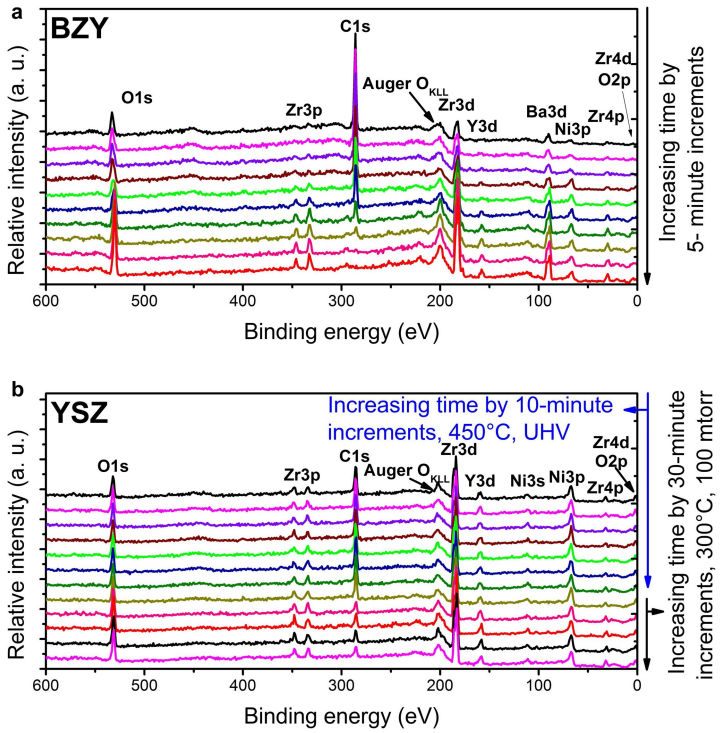
Extended Data Fig. 7 | The mechanism of SSRS and exsolution of Ni nanoparticles. a, Exposure to reducing conditions (for example, H_2 or hydrocarbon fuel) at typical PCFC operating temperatures triggers the in situ exsolution of Ni nanoparticles onto the surface of the BZY phase in the cermet anode. This exsolution process is driven by the decreased solubility of Ni in the BZY phase under anode operating conditions (500–600 °C, highly reducing) compared with the much higher starting solubility of Ni in the BZY phase under anode synthesis conditions (1,450 °C, highly oxidizing). In addition, fuel cell operation also drives the decomposition of a residual BaY_2NiO_5 phase which is a by-product of the solid-state reactive-sintering method (SSRS) used to synthesize the anode. As illustrated here, the SSRS process enables the anode to be rapidly and inexpensively fabricated in a single step starting from a homogeneous mixture of $BaCO_3$, ZrO_2, Y_2O_3, NiO and pore former, which is fired under oxidizing conditions at 1,450 °C for 18 h. During firing, a complex phase-formation and sintering process, involving the transient formation and decomposition of BaY_2NiO_5, leads to the creation of a porous two-phase NiO + Ni-doped BZY composite anode with a small amount of dispersed residual BaY_2NiO_5. During initial fuel cell operation, the NiO phase is reduced to Ni metal, while the Ni-doped BZY is reduced to BZY[36], concomitant with the exsolution of Ni nanoparticles which we hypothesize enhance the performance and durability of the fuel cell. **b**, Amplified main peaks of XRD patterns of BZY20 phase after sintering in air with different NiO amount in precursors. The amount of Ni diffusing into the BZY is determined by the defect reaction equilibrium. The amplified main XRD peaks of BZY20 phase shift to the higher angles with increasing amounts of Ni in the anode. This is consistent with a decreasing lattice constant with increasing Ni substitution on the B-site of the BZY lattice owing to the smaller size of Ni^{2+} compared with Zr^{4+} or Y^{3+}. **c**, XRD patterns of BZY20 anode after sintering in air and reduction in hydrogen at 600 °C for 150 h. **d**, Amplified main peaks of XRD patterns of BZY20 phase after sintering in air and reduction in hydrogen at 600 °C for 150 h. The peak shifts to smaller angles after reduction which is consistent with the exsolution of Ni from BZY lattice, leading to a concomitant increase in the lattice constant of the BZY phase. **e**, BaY_2NiO_5 phase formation temperature profile investigated by in situ high-temperature XRD. The colour indicates the intensity of the main peak of BaY_2NiO_5.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 8 | SEM images of the Ni nanoparticle exsolution process in the PCFC anode. a, **b**, BZY20 phase of the anode before reduction. The surface of the BZY20 phase is very clean, and there are no nanoparticles along the grain boundaries or on the triple junction points. **c**, **d**, BZY20 phase of the anode after reduction under hydrogen at 600 °C for 50 h. Nickel nanoparticles begin to form on the triple junction points and grain boundaries. **e**, **f**, BZY20 phase of the anode after reduction under hydrogen at 600 °C for 100 h. More Ni nanoparticles begin to form along grain surfaces and boundaries. The size of exsolved Ni nanoparticles is less than 100 nm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Evolution of the AP-XPS elemental survey spectrum of Ni/$BaZr_{0.9}Y_{0.1}O_{3-\delta}$ (BZY) and Ni/$Zr_{0.92}Y_{0.08}O_{2-\delta}$ (8YSZ) patterned microelectrodes. a, Ni/BZY-based microelectrode, **b**, Ni/YSZ-based microelectrode.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Performance comparison of SOFCs with PCFCs

a

Composition	T (°C)	Hydrocarbon Fuels or Alcohol	Stability (ITV=initial terminal voltage)	Degradation rate	Notes	Ref.
BZY20+Ni | BZY20 | BCFZY	500	20 vol. % C_3H_8 + 20 vol. % O_2 + 60 vol. % H_2O	j=0.2 A/cm^2	~1.5%/ 1000 hours	Cost-effective PCFC	This work
BZY20+Ni | BZY20 | BCFZY	500	10% C_4H_{10} + 20% O_2 + 70% H_2O	j=0.2 A/cm^2	No degradation after 3000 hours	Cost-effective PCFC	This work
BZY20+Ni | BZY20 | BCFZY	600	10 vol. % iso-C_4H_{10} + 20 vol. % O_2 + 70 vol. % H_2O	j_1=0.1 A/cm^2 j_2=0.2 A/cm^2 j_3=0.4 A/cm^2	~4%/ 1000 hours	Cost-effective PCFC	This work
BZY20+Ni | BZY20 | BCFZY	500	Methanol	j=0.4 A/cm^2	No degradation 8000 hours	Cost-effective PCFC	This work
BZY20+Ni | BZY20 | BCFZY	550	Ethanol	j=0.2 A/cm^2	~1.5%/ 1000hours	Cost-effective PCFC	This work
PBMO+Co-Fe/LDC/LSGM/NBSCF50-GDC	700	C_3H_8 (3%H_2O)	j=0.2 A/cm^2 ITV=0.68V	~0.6%/500 hours (~1.2%/kh)	Ni-free anode	37
RP-PSFN-CFA/LDC/LSGM/BCFN	800	C_3H_8 (3%H_2O)	j=0.4 A/cm^2 ITV=0.683V	~3.4%/100 hours (~34%/kh)	Ni-free anode	38
LSV+SDC/SDC/$La_{0.8}Sr_{0.2}CoO_{3-\delta}$ (LSC)-SDC	600	CH_4 (3%H_2O)	j=0.096 A/cm^2 ITV=0.73V	~2.6%/46 hours (~57%/kh)	Ni-free anode	39
Sn/Ni–GDC/GDC/LSCF	650	CH_4	j=0.5 A/cm^2 ITV=0.48V	No degradation 200 hours	Sn modified anode	40
Ni–YSZ(SDC)|YSZ|YSZ+LSM	800	CH_4 (3%H_2O)	j=0.33 A/cm^2 ITV=0.42V	No degradation 450 hours	SDC modified anode	41
Ni–SDC/$Ce_{0.8}Ni_{0.2}O_{2-y}$ catalytic layer/SDC/PBCO–SDC	650	CH_4 (3%H_2O)	OCV	~18%/140 hours	$Ce_{0.8}Ni_{0.2}O_{2-y}$ catalytic layer	42
Ni–SDC/$Ce_{0.8}Ni_{0.2}O_{2-y}$ catalytic layer/SDC/PBCO–SDC	650	Dry ethanol carried by N_2	OCV	~0.7%/150 hours	$Ce_{0.8}Ni_{0.2}O_{2-y}$ catalytic layer	42

b

Anode materials	Operating temperature	Fuel	Degradation rate	Comments	Reference
YSZ+Ni with Ru-CeO_2 catalyst layer	770 °C	*Iso*-octane/CO_2/steam /air mixtures	~0.9%/50 hours	Noble metal catalyst High operating temperatures	43
$Sr_2MgMoO_{6-\delta}$	800 °C	Methane	No stability on methane provided	High operating temperatures	44
BZCYYb	750 °C	Wet propane	~0.7%/24 hours	High operating temperatures, poor chemical stabilities	7
PBMO+Co-Fe catalyst	700 °C	Humidified (3% H2O) propane	~0.6%/500 hours	High operating temperatures	37
$(Pr_{0.4}Sr_{0.6})_3(Fe_{0.85}Nb_{0.15})_2O_7$ (RP-PSFN) + CoFe bimetallic alloy (CFA) nanoparticles	800 °C	Humidified (3% H_2O) propane	~3.1%/110 hours	High operating temperatures	38
Ca doped $PrBaMn_2O_{5+d}$ (PBCMO)	700 °C	Humidified (3% H_2O) propane	~3.5%/150 hours	High operating temperatures	45
YSZ+Ni+SDC catalyst layer	800 °C	Humidified (3% H_2O) methane	No degradation after ~400 hours operation	High operating temperatures	41
CMF/LSMF anode containing 10 wt% added RuO_2	800 °C	Humidified (3% H_2O) propane	~4.0%/52 hours	High operating temperatures, noble metal catalysts	46
YSZ+Ni+Ag	800 °C	Dry methane	~2.3%/100 hours	High operating temperatures	47
Double-layered perovskite $(Pr_{0.4}Sr_{0.6})_3(Fe_{0.85}Mo_{0.15})_2O_7$ (DLP-PSFM)	750 °C	Ethane	~1.5%/100 hours	High operating temperatures	48
Copper–iron–ceria–YSZ	800 °C	Methane	~21.5%/46 hours	High operating temperatures	49
$La_{0.3}Sr_{0.6}Ce_{0.1}Ni_{0.1}Ti_{0.9}O_{3-\delta}$	900 °C	Dry methane	No degradation after 80 hours	High operating temperatures	50
2.5wt.%MgO-modified Ni-SDC	800 °C	Humidified (3% H_2O) methane	~13.3%/330 hours	High operating temperatures	51
Ni– MnO/Ni–YSZ	800 °C	Humidified (3% H_2O) methane	~20.4%/24 hours	High operating temperatures	52

a, Comparison of stabilities of PCFCs and SOFCs on hydrocarbon/alcohol fuels from 2013 to 2017 (this work, and refs. [37–42]); **b**, Summary of anode materials developed for direct hydrocarbon SOFCs[7,37,38,43–52].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0054-x

Rapid energy-efficient manufacturing of polymers and composites via frontal polymerization

Ian D. Robertson[1,2,7], Mostafa Yourdkhani[1,7], Polette J. Centellas[1,3], Jia En Aw[1,3], Douglas G. Ivanoff[1,4], Elyas Goli[1,5], Evan M. Lloyd[1,6], Leon M. Dean[1,4], Nancy R. Sottos[1,4], Philippe H. Geubelle[1,3], Jeffrey S. Moore[1,2] & Scott R. White[1,3]*

Thermoset polymers and composite materials are integral to today's aerospace, automotive, marine and energy industries and will be vital to the next generation of lightweight, energy-efficient structures in these enterprises, owing to their excellent specific stiffness and strength, thermal stability and chemical resistance[1–5]. The manufacture of high-performance thermoset components requires the monomer to be cured at high temperatures (around 180 °C) for several hours, under a combined external pressure and internal vacuum[6]. Curing is generally accomplished using large autoclaves or ovens that scale in size with the component. Hence this traditional curing approach is slow, requires a large amount of energy and involves substantial capital investment[6,7]. Frontal polymerization is a promising alternative curing strategy, in which a self-propagating exothermic reaction wave transforms liquid monomers to fully cured polymers. We report here the frontal polymerization of a high-performance thermoset polymer that allows the rapid fabrication of parts with microscale features, three-dimensional printed structures and carbon-fibre-reinforced polymer composites. Precise control of the polymerization kinetics at both ambient and elevated temperatures allows stable monomer solutions to transform into fully cured polymers within seconds, reducing energy requirements and cure times by several orders of magnitude compared with conventional oven or autoclave curing approaches. The resulting polymer and composite parts possess similar mechanical properties to those cured conventionally. This curing strategy greatly improves the efficiency of manufacturing of high-performance polymers and composites, and is widely applicable to many industries.

Present technologies for manufacturing high-performance thermoset and fibre-reinforced polymer composite (FRPC) parts rely on curing in large, expensive autoclaves or ovens. For example, curing a small section of the Boeing 787's carbon fibre/epoxy fuselage is estimated to require 350 gigajoules (GJ) of energy during its eight-hour cure cycle, producing more than 80 tons of carbon dioxide[7]. Consequently, there has been much interest in producing these materials with less energy, reducing their cost and environmental impact and furthering their application in commercial markets[6,8–11]. Frontal polymerization is a promising curing strategy that substantially reduces manufacturing burdens by using the enthalpy of polymerization to provide the energy for materials synthesis, rather than requiring an external energy source[12]. In frontal polymerization, a solution of a monomer and a latent initiator is heated locally until the initiator is activated for polymerization of the monomer, producing heat from the polymerization that further drives the reaction. The autoactivation process produces a propagating reaction wave that rapidly transforms the available monomer into polymer. Frontal polymerization has been used to synthesize a variety of polymeric materials, including functionally graded polymers, nanocomposites, hydrogels, sensory materials and FRPCs[13–21]. Most of the materials used in frontal polymerization to date, however, are unsuitable for high-performance applications. For example, although acrylate monomers possess the requisite energy density and reactivity to frontally polymerize, the mechanical properties of the resulting polymers are greatly inferior to those used in structural FRPCs. By contrast, epoxy monomers produce mechanically robust polymers, but are challenging to frontally polymerize because of their lower reactivity[20,22]. Moreover, it is essential for successful manufacturing of FRPCs that the liquid monomer be stable and essentially free of background polymerization at room temperature. These requirements motivate the development of frontal-polymerization chemistry with a controllable and stable processing window, a high energy density and reactivity, and a mechanically and thermally robust polymer product.

Here, we demonstrate that well-controlled frontal polymerization facilitates the rapid production of high-performance thermoset and FRPC parts with minimal energy input. Furthermore, the process is compatible with commonly used manufacturing techniques and produces high-quality thermoset materials. Frontal curing of FRPCs is challenging because a high volume fraction of fibres is necessary to produce a composite material with good mechanical properties, yet the corresponding reduction in resin content reduces the exothermic energy density available for frontal polymerization. As such, the frontal-polymerization chemistry must have a high molar enthalpy of polymerization and sufficiently high rate of polymerization to prevent frontal quenching. Fabricating small components with frontal polymerization is similarly challenging because much of the heat of polymerization is lost to the environment through air or tooling surfaces[23,24].

The frontal ring-opening metathesis polymerization (FROMP) of dicyclopentadiene (DCPD) using a thermally activated ruthenium catalyst exhibits the high energy density, high reactivity and low viscosity required for the synthesis of high-performance thermosets (Fig. 1a). The resulting polydicyclopentadiene (pDCPD) is a cross-linked thermoset polymer that is suitable for the fabrication of durable resin and FRPC parts, owing to its high fracture toughness, impact resistance, stiffness and chemical resistance[25–27]. However, FROMP chemistry has been severely limited in the past by its short pot life of less than 30 minutes[28,29]. Recently, we demonstrated that inhibitors of the alkyl phosphite family substantially extend the room-temperature liquid-processing window for FROMP of DCPD up to 30 hours[30]. Here, we use phosphite-inhibited FROMP of DCPD to efficiently fabricate neat pDCPD and carbon FRPC structures. Compared with conventional curing, our frontal-polymerization strategy reduces energy requirements by more than ten orders of magnitude for large components (Fig. 1b).

Phosphite-inhibited DCPD containing second-generation Grubbs' catalyst (GC2) slowly transforms at room temperature from a liquid to a viscoelastic gel. Remarkably, the gelation of the monomer does not result in concomitant spontaneous polymerization, as observed with previous FROMP chemistries. Tuning the inhibitor concentration allows access to a range of rheological profiles between the low-viscosity

[1]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [2]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [3]Department of Aerospace Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [4]Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [5]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [6]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [7]These authors contributed equally: Ian D. Robertson, Mostafa Yourdkhani. *e-mail: swhite@illinois.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

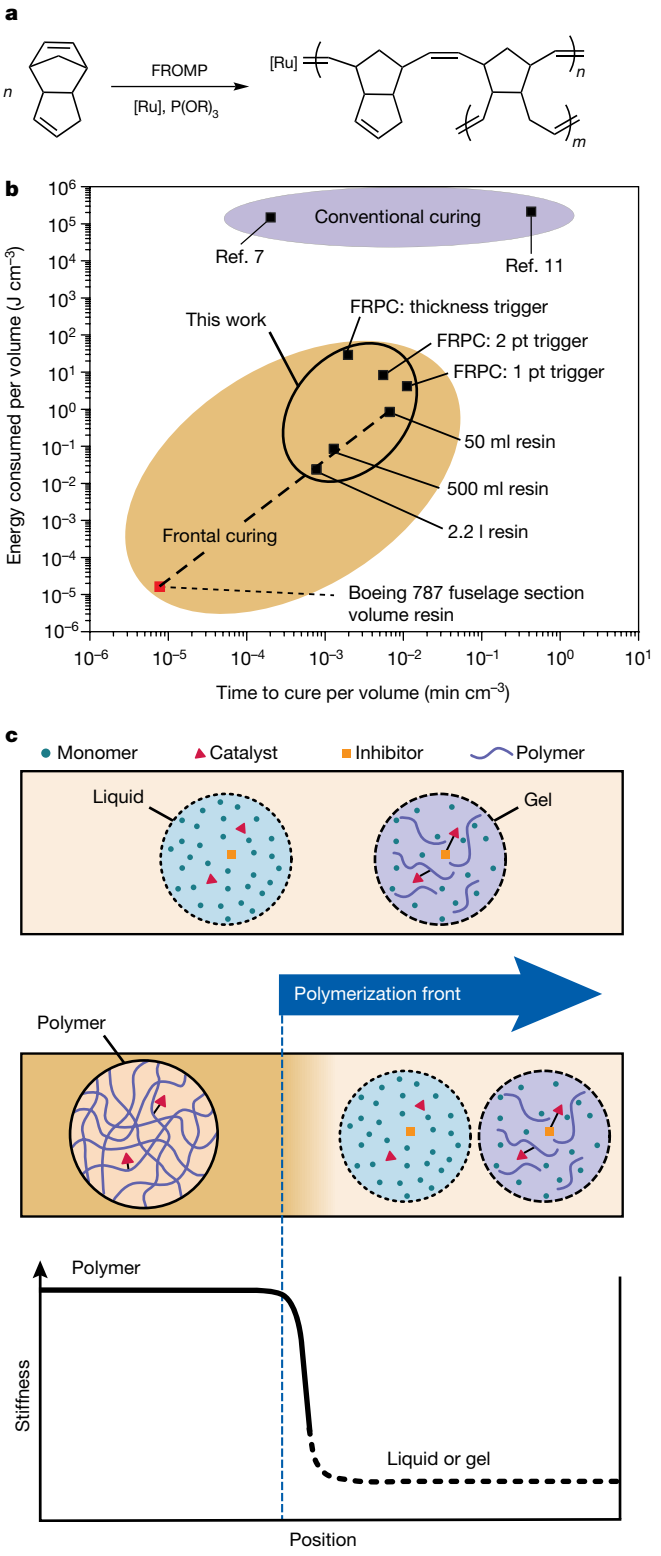


Fig. 1 | Overview of the frontal-polymerization concept. a, Scheme for the FROMP of DCPD using a ruthenium catalyst (Ru; that is, second-generation Grubbs' catalyst) and an alkyl phosphite inhibitor ($P(OR)_3$). **b**, Comparison of the energy required for curing versus the time taken for curing via frontal polymerization and conventional curing. Given that the only energy needed for frontal curing is the initiating stimulus, the process requires several orders of magnitude less energy than conventional curing, while also occurring faster and with less expensive equipment. References 7 and 11 describe the manufacturing of a section of the Boeing 787 fuselage and a 900 cm^2 carbon FRPC panel, respectively. We calculated that the curing of a section of the Boeing 787 fuselage by FROMP (red square) reduces the energy consumption by ten orders of magnitude compared with conventional techniques. **c**, In the technique used here, the FROMP solution is triggered to polymerize in its liquid stage, or allowed to form a gel at room temperature and later activated. In both cases, a rapid FROMP reaction transforms the liquid or gel into a durable thermoset polymer.

liquid and free-standing elastomeric gel, all of which frontally polymerize upon thermal activation (Fig. 1c). These free-standing gels are deformable and easily embossed; the resulting patterned gel structure can be quickly fixed into a rigid pDCPD structure via FROMP (Fig. 2a–c and Supplementary Video 1). Moreover, when the liquid monomer is poured onto a micropatterned substrate, peeled off in the gel stage, and polymerized by FROMP, a high-fidelity replica is produced in one step (Fig. 2d, e).

The gel is amenable to three-dimensional (3D) printing during the high-viscosity fluid stage, whereby the viscous liquid is extruded from a print head and frontally polymerized immediately upon exiting the nozzle, allowing for the simultaneous free-form printing and curing of thermoset polymers (Fig. 2f and Supplementary Video 2). Matching the print-head velocity to the front velocity allows for the manufacture of free-form complex architectures that are not possible with traditional additive manufacturing approaches (Fig. 2g, h). Once printing is complete, the part is fully cured (Extended Data Fig. 1c) and there is no need for further processing.

In marked contrast with traditional autoclave processing, we can fabricate FRPC parts in less than 5 minutes through FROMP of woven carbon fibres infused with the monomer solution (Fig. 2i, j). The very low viscosity of the liquid monomer at room temperature (~1.5 cP) allows the resin to be rapidly infused into continuous-fibre layups that have a high volume fraction of fibres via out-of-autoclave processing techniques such as vacuum-assisted resin-transfer moulding (VARTM). FROMP is triggered by briefly powering a resistive heating wire embedded in the layup, which provides sufficient thermoelectric stimulus (Fig. 3a and Supplementary Video 3). The high reactivity of DCPD facilitates FROMP in thin FRPC laminates with up to 50 vol% fibre reinforcement, while remarkably reducing the manufacturing time and required energy. We can further reduce the manufacturing time by using multiple triggering points (Fig. 3b and Supplementary Video 4) or by propagating the reaction through the thickness using a resistive heater underneath the layup (Fig. 3c and Supplementary Video 5). However, commensurately more triggering energy is required in these configurations (Fig. 3a–c). We expected the relatively low volume fraction of monomers in the FRPCs to reduce the frontal velocity. Surprisingly, however, the frontal velocity in the FROMP-FRPC (9.8 cm min^{-1}) was higher than that observed in neat resin for the same formulation (7.5 cm min^{-1}). We surmise that the carbon fibre conducts heat from the exothermic reaction and preheats a region of monomer ahead of the front, accelerating the reaction and producing a higher frontal velocity.

Numerical simulation of a carbon-fibre bundle, or tow, in the resin supports this hypothesis. Using a simple model system—which consists of a single carbon-fibre tow made of 3,000 individual fibres suspended in neat resin—we observe that the carbon fibres accelerate FROMP and produce a distinct change in the front shape from flat to conical (Fig. 3d, e). To provide analytical insight, we carried out numerical simulations of the thermochemical problem using a transient, nonlinear finite-element solver. We used an adaptive meshing scheme to capture the sharp gradients in temperature and degree of cure in the immediate vicinity of the propagating polymerization front. As shown by the computed thermal solutions in Fig. 3d, e, the higher thermal conductivity of the fibre tow changes the thermal field ahead of the front, thereby modifying the shape and speed of the front. The numerical solver provides useful insights into FROMP by creating a direct link between the cure kinetics model and FROMP propagation. For example, simulations of multipoint initiation predict a temperature spike upon front convergence, because the heat of reaction is no longer

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

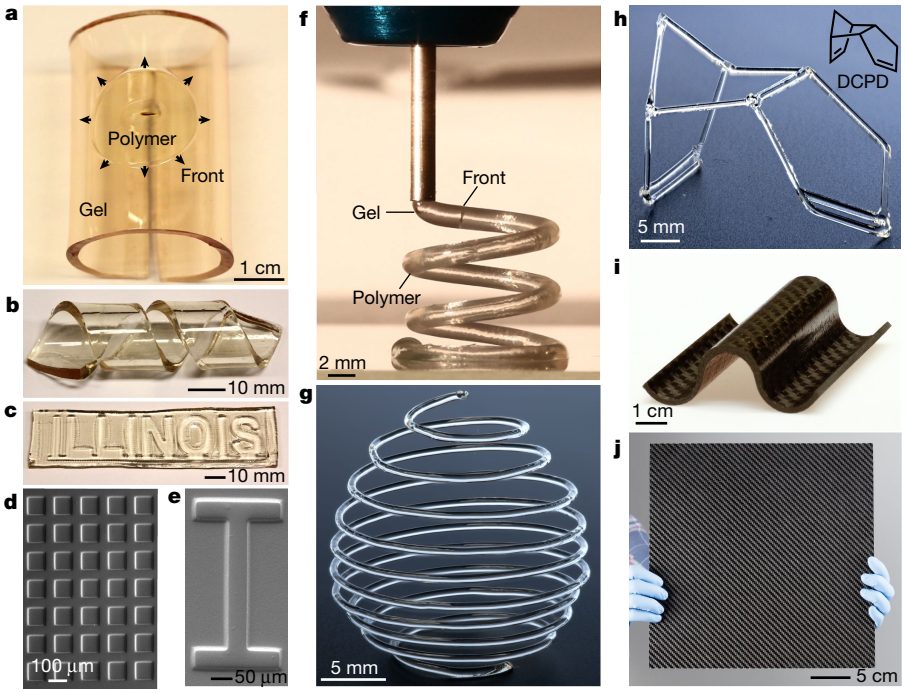


Fig. 2 | Advanced manufacturing with FROMP. a, FROMP in a free-standing gel, propagating radially from a single initiation-point source. **b**–**e**, Macropatterned/micropatterned pDCPD produced by gel FROMP. In **b**, a flat gel sheet is rolled into a helix structure before FROMP. In **c**, a flat gel sheet is imprinted with an ILLINOIS stamp before FROMP. In **d** and **e**, gel sheets are produced by moulding at room temperature for 18 hours and followed by FROMP to solidify the micropattern. **f**, 3D printing of gel DCPD solution that is solidified by FROMP immediately following extrusion from the print head. **g**, **h**, Free-form 3D-printed structures produced via FROMP. **i**, A corrugated carbon FRPC part fabricated by FROMP using vacuum-assisted resin-transfer moulding. **j**, A 900 cm^2 carbon FRPC panel (with 51% fibre volume fraction) cured by FROMP in 5 minutes using about 750 J of energy.

conducted away from the front. This thermal overshoot is observed in two-point initiation experiments (Fig. 3b), and, if excessive, may lead to material degradation near the location of merging fronts. Optimization strategies for the manufacture of FRPCs with minimum curing time and required energy coupled with high quality will require similar computational modelling approaches.

Carefully selecting the phosphite inhibitor and its concentration enables us to control room-temperature rheological profiles as well as frontal velocity (Fig. 4a). Figure 4b divides the rheological characteristics of a particular formulation into three regimes. Prior to the gel point, liquid FROMP is suitable for a wide variety of infusion techniques that require a low-viscosity resin. After the gel point, the material exhibits increased

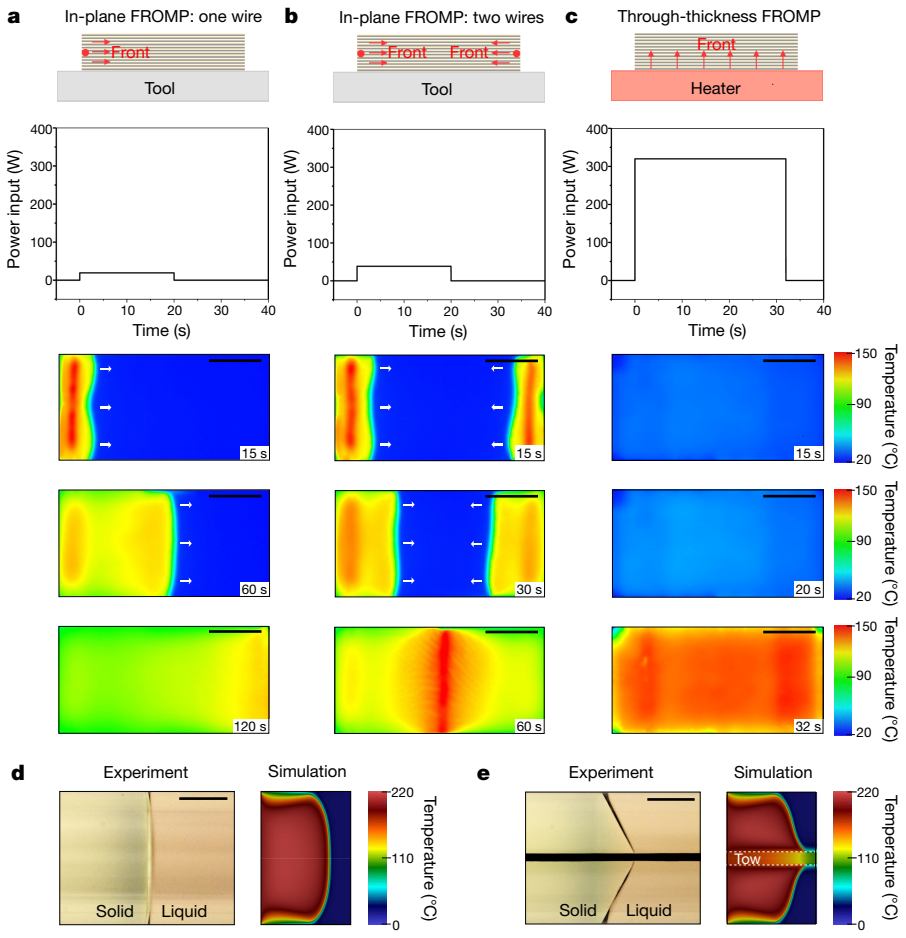


Fig. 3 | Composite fabrication and modelling. The FROMP curing of FRPC is a versatile process. **a**, Heating one embedded wire ignites a single front that propagates (left to right) to fully cure the FRPC in about 2 minutes. **b**, Heating two embedded wires (opposite ends) ignites two fronts that merge in the middle, completing curing in half the time (about 1 minute). **c**, Heating the layup from below ignites a front that travels through the thickness of the panel and completes the cure in about 30 seconds. Still images are captured using a thermal infrared camera. Scale bar in **a**–**c** represents 5 cm. **d**, A propagating FROMP reaction wave captured experimentally and via simulation. The optical difference between the solid and liquid phases shows that the front shape is slightly convex in the direction of propagation (left to right). This frontal shape is closely matched by simulation. **e**, In the presence of a single carbon-fibre tow, the frontal shape is altered and the front is accelerated near the tow surface. The increase in front speed from **d** to **e** is roughly 10%. Scale bar in **d**, **e** represents 4 mm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

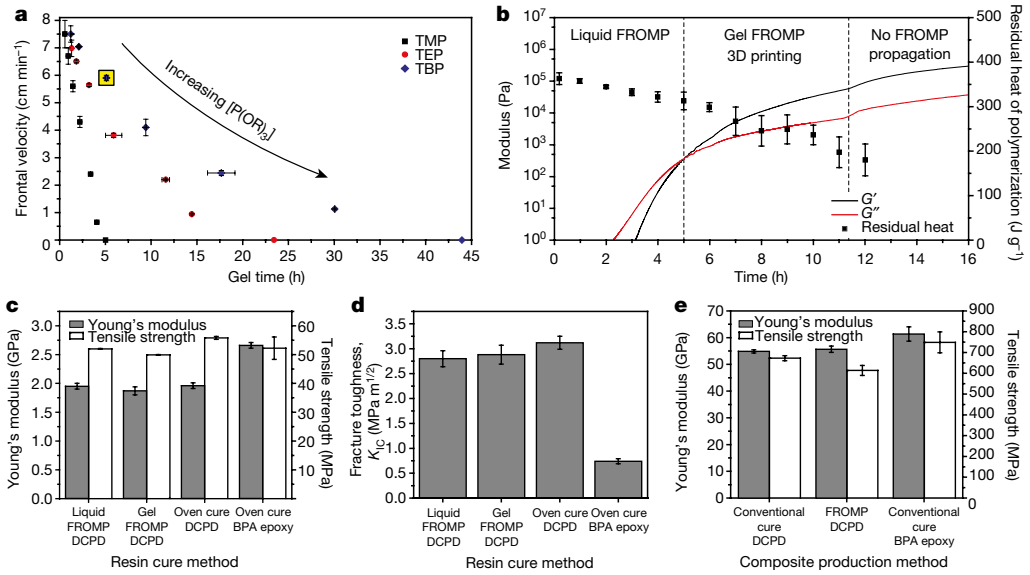


Fig. 4 | Characterization of the FROMP curing approach. a, Gel time and frontal velocity are controlled by the identity of the phosphite inhibitor (TMP, trimethyl phosphite; TEP, triethyl phosphite; TBP, tributyl phosphite) and its concentration (0.3, 0.5, 1, 2, 4 or 8 molar equivalents with respect to GC2 catalyst). FROMP is no longer observed with 9, 12 or 14 molar equivalents of TMP, TEP and TBP, respectively. Error bars represent standard deviation from the mean ($n = 3$). **b**, Representative rheological and thermal behaviour for curing at room temperature (23 °C) of a FROMP solution containing 1 molar equivalent TBP inhibitor with respect to GC2 catalyst (highlighted yellow data point in **a**), showing three rheological regimes. The liquid FROMP stage is suitable for resin-infusion/transfer manufacturing. After gelation, both 3D printing and the macropatterning/micropatterning of gels are possible. After vitrification, FROMP is no longer possible. The residual heat at each stage is determined by DSC analysis. The initial data point at $t = 0.19$ hours indicates the total enthalpy of the reaction of dicyclopentadiene ($H_r = 362.8 \pm 13.1$ J g^{-1}). Error bars represent standard deviation from the mean ($n = 3$). G', storage modulus; G'', loss modulus. **c**, Young's modulus and tensile strength for liquid FROMP, gel FROMP, and oven-cured pDCPD. A bisphenol A (BPA) aerospace-grade epoxy is shown for comparison. Error bars represent standard deviation from the mean ($n = 7$). **d**, Mode I fracture toughness (K_{IC}, where K is the stress intensity factor, I denotes 'mode I crack opening', and C denotes 'critical') for liquid FROMP ($n = 5$), gel FROMP ($n = 5$) and oven-cured pDCPD ($n = 6$) in comparison with BPA aerospace-grade epoxy ($n = 8$). Error bars represent standard deviation from the mean. **e**, Young's modulus and tensile strength for conventional-cured (51% volume fraction; $n = 10$) and FROMP-cured (51% volume fraction; $n = 13$) pDCPD-FRPCs in comparison with aerospace-grade (BPA) carbon/epoxy FRPC (52% volume fraction; $n = 8$). Error bars represent standard deviation from the mean.

elasticity, enabling access to additional processing techniques. During the first part of the post-gelation regime, the monomer solution is a viscous liquid suitable for 3D printing (vide supra). It exhibits shear thinning behaviour (Extended Data Fig. 1d) and is readily extruded from a print head under pressure. In the latter stages of this regime, the material forms a free-standing gel that is resistant to plastic flow. This stage is useful for casting micropatterns and imprinting (vide supra). The regime ends with a secondary vitrification transition that corresponds to the point at which there is no longer sufficient chemical energy in the reactive formulation to sustain FROMP. For larger concentrations of tributyl phosphite (TBP), this transition occurs more than 40 hours after mixing at 23 °C, providing sufficient time for manufacturing. For stiff gel applications, a trimethyl phosphite (TMP) formulation that gels quickly may be desirable. For 3D printing and composite manufacturing, a longer processing window is desired; therefore, TBP or triethyl phosphite (TEP) should be used at the highest concentration that still enables FROMP.

We monitored the residual exotherm by differential scanning calorimetry (DSC) during room-temperature curing, to examine the effect of gelation on the cure profile and to correlate the rheological transitions with the degree of cure (Fig. 4b). We found that there is a slight shift of the exothermic peak to lower temperatures as the curing proceeds, and that the peak broadens substantially (Extended Data Fig. 1a). This behaviour suggests that the reactive formulation becomes more sensitive to increases in temperature during gelation; this sensitivity facilitates frontal polymerization in the partially cured state. The increase in reactivity may result from the formation of new catalyst species in situ, as seen in ^{31}P nuclear magnetic resonance (NMR) spectroscopy of the gelled material[30]. The pDCPD produced after both liquid and gel FROMP exhibits a minimal residual exotherm, indicating that a complete cure is achieved (Extended Data Fig. 1b).

FROMP-cured pDCPD exhibits comparable mechanical properties to conventionally (oven) cured pDCPD. The tensile strength and stiffness of pDCPD that has been frontally polymerized from both the liquid and the gel stage indicate that frontal polymerization provides the energy necessary for complete cross-linking of the monomer (Fig. 4c). Comparing the tensile properties of pDCPD with those of an aerospace-grade bisphenol A (BPA) epoxy system indicates that this new curing strategy and matrix resin produce polymers suitable for use in high-performance applications. In fact, the enhanced fracture toughness of FROMP-pDCPD compared with BPA epoxy (Fig. 4d) may provide high-performance composites with superior energy absorption and fatigue resistance[27].

Frontally polymerized FRPCs also exhibit mechanical properties comparable to those of FRPCs manufactured using a conventional aerospace-grade BPA epoxy system (Fig. 4e). The carbon fibre used in our FRPCs was surface-treated to allow interaction with epoxy functional groups; we envisage that greater tensile properties for pDCPD-FRPCs should be made possible by tailoring the interfacial interaction of the carbon fibre and pDCPD through proper fibre surface treatments. We attribute the slightly lower tensile strength of FROMP-cured compared with conventionally cured pDCPD composite specimens to the lower degree of cross-linking, resulting from heat loss through the fibre reinforcement and tooling (Extended Data Table 2).

In conclusion, we have developed and demonstrated a new frontal polymerization curing method for the rapid, energy-efficient manufacturing of thermoset pDCPD and FRPCs. This method is applicable to moulding, imprinting, 3D printing and resin-infusion techniques, and the resulting polymers and carbon-fibre-reinforced polymer composites exhibit excellent mechanical properties, comparable to those of materials produced by conventional techniques. We envisage that the controlled frontal polymerization of high-quality resins will

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

enable a variety of new manufacturing technologies—such as on-site, on-demand manufacturing, in-the-field repair of FRPCs and mould-less production—owing to the spatiotemporal control over the curing process.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0054-x.

Received: 5 December 2017; Accepted: 19 February 2018;
Published online 9 May 2018.

1. Pascault, J.-P., Sautereau, H., Verdu, J. & Williams, R. J. J. *Thermosetting Polymers* (CRC Press, New York, 2002).
2. Daniel, I. M. & Ishai, O. *Engineering Mechanics of Composite Materials* (Oxford Univ. Press, Oxford, 2005).
3. Brøndsted, P., Lilholt, H. & Lystrup, A. Composite materials for wind power turbine blades. *Annu. Rev. Mater. Res.* **35**, 505–538 (2005).
4. Friedrich, K. & Almajid, A. A. Manufacturing aspects of advanced polymer composites for automotive applications. *Appl. Compos. Mater.* **20**, 107–128 (2013).
5. Karbhari, V. M. & Seible, F. Fiber reinforced composites—advanced materials for the renewal of civil infrastructure. *Appl. Compos. Mater.* **7**, 95–124 (2000).
6. Abliz, D. et al. Curing methods for advanced polymer composites—a review. *Polym. Polymer Compos.* **21**, 341–348 (2013).
7. Timmis, A. J. et al. Environmental impact assessment of aviation emission reduction through the implementation of composite materials. *Int. J. Life Cycle Assess.* **20**, 233–243 (2015).
8. Taynton, P. et al. Repairable woven carbon fiber composites with full recyclability enabled by malleable polyimine networks. *Adv. Mater.* **28**, 2904–2909 (2016).
9. Lee, J., Stein, I. Y., Kessler, S. S. & Wardle, B. L. Aligned carbon nanotube film enables thermally induced state transformations in layered polymeric materials. *ACS Appl. Mater. Interfaces* **7**, 8900–8905 (2015).
10. Li, N., Li, Y., Hang, X. & Gao, J. Analysis and optimization of temperature distribution in carbon fiber reinforced composite materials during microwave curing process. *J. Mater. Process. Technol.* **214**, 544–550 (2014).
11. Witik, R. A. et al. Economic and environmental assessment of alternative production methods for composite aircraft components. *J. Clean. Prod.* **29–30**, 91–102 (2012).
12. Pojman, J. A. in *Polymer Science: A Comprehensive Reference* Vol. 4 (eds. Matyjaszewski, K. & Möller, M.) 957–980 (Elsevier, Amsterdam, 2012).
13. Chekanov, Y. A. & Pojman, J. A. Preparation of functionally gradient materials via frontal polymerization. *J. Appl. Polym. Sci.* **78**, 2398–2404 (2000).
14. Nuvoli, D. et al. Synthesis and characterization of functionally gradient materials obtained by frontal polymerization. *ACS Appl. Mater. Interfaces* **7**, 3600–3606 (2015).
15. Chen, S., Sui, J., Chen, L. & Pojman, J. A. Polyurethane–nanosilica hybrid nanocomposites synthesized by frontal polymerization. *J. Polym. Sci. A* **43**, 1670–1680 (2005).
16. Yan, Q.-Z., Zhang, W.-F., Lu, G.-D., Su, X.-T. & Ge, C.-C. Frontal polymerization synthesis of starch-grafted hydrogels: effect of temperature and tube size on propagating front and properties of hydrogels. *Chem. Eur. J.* **12**, 3303–3309 (2006).
17. Guo, X., Wang, C.-F., Fang, Y., Chen, L. & Chen, S. Fast synthesis of versatile nanocrystal-embedded hydrogels toward the sensing of heavy metal ions and organoamines. *J. Mater. Chem.* **21**, 1124–1129 (2011).
18. Sanna, R. et al. Polymer hydrogels of 2-hydroxyethyl acrylate and acrylic acid obtained by frontal polymerization. *J. Polym. Sci. A* **50**, 1515–1520 (2012).
19. Nagy, I. P., Sike, L. & Pojman, J. A. Thermochromic composite prepared via a propagating polymerization front. *J. Am. Chem. Soc.* **117**, 3611–3612 (1995).
20. White, S. R. & Kim, C. A simultaneous lay-up and in situ cure process for thick composites. *J. Reinf. Plast. Compos.* **12**, 520–535 (1993).
21. Kim, C., Teng, H., Tucker, C. L. & White, S. R. The continuous cuing process for thermoset polymer composites. Part 1: modeling and demonstration. *J. Compos. Mater.* **29**, 1222–1253 (1995).
22. Scognamillo, S., Bounds, C., Luger, M., Mariani, A. & Pojman, J. A. Frontal cationic curing of epoxy resins. *J. Polym. Sci. A* **48**, 2000–2005 (2010).
23. Robertson, I. D., Lopez Hernandez, H., White, S. R. & Moore, S. Rapid stiffening of a microfluidic endoskeleton via frontal polymerization. *ACS Appl. Mater. Interfaces* **6**, 18469–18474 (2014).
24. Datta, P., Efimenko, K. & Genzer, J. The effect of confinement on thermal frontal polymerization. *Polym. Chem.* **3**, 3243–3246 (2012).
25. Delaude, L. & Noels, A. F. in *Kirk-Othmer Encyclopedia of Chemical Technology* Vol. 26 (ed. Seidel, A.) 920–958 (John Wiley & Sons, Inc., Malden, 2007).
26. Woodson, C. S. Jr & Grubbs, R. H. Polymeric composites including dicyclopentadiene and related monomers. US patent 6,310,121 B1 (2001).
27. Vallons, K. A. M., Drozdzak, R., Charret, M., Lomov, S. V. & Verpoest, I. Assessment of the mechanical behaviour of glass fibre composites with a tough polydicyclopentadiene (PDCPD) matrix. *Compos. Part A Appl. Sci. Manuf.* **78**, 191–200 (2015).
28. Mariani, A., Fiori, S., Chekanov, Y. & Pojman, J. A. Frontal ring-opening metathesis polymerization of dicyclopentadiene. *Macromolecules* **34**, 6539–6541 (2001).
29. Ruiu, A., Sanna, D., Alzari, V., Nuvoli, D. & Mariani, A. Advances in the frontal ring opening metathesis polymerization of dicyclopentadiene. *J. Polym. Sci. A* **52**, 2776–2780 (2014).
30. Robertson, I. D. et al. Alkyl phosphite inhibitors for frontal ring-opening metathesis polymerization greatly increase pot life. *ACS Macro Lett.* **6**, 609–612 (2017).

Acknowledgements This research was conducted as part of the Center for Excellence for Self-Healing, Regeneration and Structural Remodeling, supported by the United States Air Force Office of Scientific Research through award FA9550-16-1-0017. We thank J. Sung for preparing the micropatterned silicon substrates, T. Ross for sample photography, D. Loudermilk for graphics assistance, and the Beckman Institute for Advanced Science and Technology for use of their facilities and equipment. I.D.R. thanks the US Department of Defense for a National Defense Science and Engineering Graduate Fellowship. L.M.D. thanks the National Science Foundation for a Graduate Research Fellowship.

Reviewer information *Nature* thanks J. Pojman and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.R.W., J.S.M., N.R.S. and P.H.G. directed the research. J.S.M., S.R.W., N.R.S. and I.D.R. conceived the idea. I.D.R., M.Y., P.J.C., J.E.A., D.G.I., E.M.L. and L.M.D. performed the experiments. E.G. conducted the computational studies. All authors participated in writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0054-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0054-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.R.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Materials. Dicyclopentadiene (DCPD), 5-ethylidene-2-norbornene (ENB), second-generation Grubbs' catalyst (GC2), phenylcyclohexane, and phosphite inhibitors (TMP, TEP and TBP) were purchased from Sigma-Aldrich and used as received without further purification. For all FRPC specimens, the fibre reinforcement is Toray T300 carbon fibre 2 × 2 twill weave fabric (tow size 3,000; areal density 204 g m^{-2}). The resistive heating wire used to trigger FROMP in the manufacturing of neat resin panels and in FRPC manufacturing is a 26-gauge Kanthal wire (diameter 0.40 mm; resistivity 1.4×10^{-4} Ω cm).

Frontal polymerization. Given that DCPD is solid at room temperature, we first melt it in an oven at 35 °C and then add 5 wt.% ENB to depress the melting point. All references to DCPD herein refer to this 95/5 DCPD/ENB solution. This mixture is then degassed at 16 kPa overnight. In a typical experiment, we weigh out 3.21 mg GC2 into an Eppendorf tube and dissolve it in 400 μl phenylcyclohexane. We add an appropriate amount of phosphite inhibitor (0–14 molar equivalents with respect to GC2) to the solution via a volumetric syringe. The catalyst/inhibitor solution is then added to 5 g DCPD (10,000 molar equivalents with respect to GC2) and thoroughly mixed.

Unless otherwise specified, the solution is frontally polymerized immediately after mixing. Different types and concentrations of inhibitor are used for each manufacturing technique to tune the reaction kinetics based on the requirements of the target application. The various inhibitor concentrations and resin incubation times are summarized in Extended Data Table 1.

FROMP of stiff gel. We use TBP (2 molar equivalents with respect to GC2) as the inhibitor for the free-standing gel specimens shown in Fig. 2a–c. The solution is poured into a flat glass plate mould with a polyurethane rubber gasket and allowed to polymerize for 18 hours at room temperature. The glass plates are then carefully removed and the free-standing gel is extracted from the mould. Gels are then either deformed by hand (Fig. 2a, b) or embossed with a plastic stamp (Fig. 2c) to create the pattern. A soldering iron is used to locally heat a single point on the gel, initiating the FROMP reaction.

For micropatterned specimens (Fig. 2d, e), a patterned silicon wafer is attached to a microscope glass slide using cyanoacrylate adhesive. A second microscope slide is used with a polyurethane rubber gasket to form a small glass plate mould. The same chemistry described previously for free-standing gel experiments is used to fill the mould and the solution is allowed to polymerize for 18 hours. The gel is carefully peeled off the micropatterned surface and examined via optical microscopy to confirm pattern transfer. The gel is then suspended pattern-side down between two microscope slides and FROMP is initiated by applying a soldering iron to one end of the sample.

Front velocity and temperature measurements. We measure the peak temperature during FROMP by inserting a T-type thermocouple (TMQSS, Omega) into the liquid resin, gel or FRPC specimens before initiating FROMP. For experiments performed in cylindrical test tubes, we insert the thermocouple into the centre of the test tube to reduce heat-transfer effects at the test-tube wall. Similarly, for experiments performed in composite layups, the thermocouple is inserted at the mid-plane of the layup.

Frontal propagation is tracked with a Canon EOS 7D digital camera or an FLIR SC620 thermal infrared camera. Front velocity is extracted from the slope of the best-fit trendline for front position versus time. We identify the front position by a colour change and refractive-index mismatch in optical images, or a sharp thermal gradient in infrared images.

Rheological measurements. Isothermal rheological measurements are performed using a TA Instruments AR-G2 rheometer equipped with 25-mm-diameter parallel aluminium plates and a solvent trap. An appropriate amount of phosphite inhibitor (0–14 molar equivalents with respect to GC2) is dissolved in the monomer solution. Time-sweep measurements are performed at 23 °C with a strain of 0.1% and a frequency of 1 Hz. The gel times plotted in Fig. 4a correspond to the crossover of the storage modulus, G', and loss modulus, G'', as shown in Fig. 4b. For the shear thinning behaviour of the 3D printable DCPD ink (Extended Data Fig. 1d), viscosity and stress data are obtained in a flow sweep test with descending shear rate at a fixed temperature of −5 °C.

Heat of reaction and degree of cure analysis. We carry out DSC measurements on a TA Instruments Q20 DSC equipped with a CFL-50 cooling system. Samples are transferred into aluminium hermetic DSC pans at room temperature and sealed. The sample mass is determined using an analytical balance (XPE205, Mettler-Toledo). The mass of liquid resin samples is carefully maintained between 2 mg and 3 mg because of the highly exothermic nature of the studied reaction, as greater masses exceed the instrument's ability to maintain a consistent temperature gradient. We determine cure profiles of liquid resins and gels at temperatures between −50 °C and 250 °C with constant ramp rates. The enthalpy of reaction is determined through the integration of heat flow over the exothermic peak after baseline correction. The specific heat capacity is determined between 25 °C and 200 °C by comparison with a sapphire standard.

3D printing. A DCPD/GC2 solution containing 0.5 molar equivalents of TEP with respect to GC2 is transferred into a 3 ml syringe barrel and left to stand at 23.0 °C in an environmental test chamber (MicroClimate, Cincinnati Sub-Zero Products) for 160 minutes to allow the cross-linking of the ink mixture to form a viscous liquid. Ink in the syringe barrel is then sealed with a piston and is ready for printing. The 3D printer consists of an air-operated high pressure dispensing tool (HP 3CC, Nordson EFD) mounted on a robotic motion-controlled stage (JL2000, Robocasting Enterprises). The pressure actuation and stage motion is simultaneously controlled via custom-designed software (RoboCAD 2.0). Air is supplied from a compressed air tank (Airgas) using a customized pressure controller. The print bed consists of an aluminium plate heated by a polyimide film heater (KH-608, OMEGA) to 70 °C. Glass slides are placed on the print bed to capture the prints. The barrel housing is fitted with stainless-steel dispenser tips before printing (inner diameters between 0.25 mm and 1.55 mm; Nordson EFD). The ink barrel is inserted into the dispensing tool, then cooled down to −5 °C ± 3 °C by surrounding the dispensing tool with dry ice. A few seconds after DCPD ink is printed on the heated glass slide, the front initiates via the heated print bed. The front then propagates along the filament, following the print head during printing. The shear thinning behaviour of the gel (Extended Data Fig. 1d) is crucial for the 3D printing of free-form structures. Using DSC, we estimate the degree of cure of the unprinted ink and printed specimen to be about 23.6% and 99.2%, respectively (Extended Data Fig. 1c).

Fabrication and testing of neat resin panels. We use a resin solution containing 1 molar equivalent of TBP with respect to GC2 for the fabrication of both FROMP-pDCPD and conventionally cured pDCPD resin panels. The resin mixture is degassed for 10 minutes at 10 kPa and then poured into cell casting moulds. We use a 215.9 mm × 203.2 mm × 6.4 mm U-shaped polyurethane spacer between two glass plates for tensile testing specimens, and a 139.7 mm × 127.0 mm × 9.5 mm U-shaped polyurethane spacer between two glass plates for fracture toughness specimens. Panels are manufactured using liquid FROMP, gel FROMP, and conventional (oven) cure. Liquid FROMP panels are initiated immediately by applying an electric current to a resistive wire placed along one edge of the mould. The power supply is turned off as soon as frontal propagation is observed, and remains off for the remainder of the propagation. We keep gel FROMP panels at 23 °C for approximately 6 hours, at which time the resin becomes a rubbery gel, before initiating the FROMP via a resistive wire placed along one edge of the mould. The power supply is turned off as soon as frontal propagation is observed, and remains off for the remainder of the propagation. Conventional-cured panels are cured for 24 hours at 30 °C, 2 hours at 70 °C and 1.5 hours at 170 °C.

For comparison, we also manufacture panels of an aerospace-grade bisphenol A (BPA) epoxy resin. A solution of Araldite LY 8605 (resin) and Aradur 8605 (hardener) is mixed (100/35 weight ratio) and then degassed for 1 hour at an absolute pressure of 10 kPa. The epoxy system is poured into cell casting moulds similar to those used for the pDCPD panels. Samples are cured according to the manufacturer's recommended cure cycle, which is 24 hours at room temperature, 2 hours at 121 °C and 3 hours at 177 °C.

We cut dog-bone specimens for tensile testing from the manufactured panels following ASTM standard D638 type I dimensions. Tensile tests are conducted using an Instron 5984 with a 150 kN load cell and a video extensometer is used to measure the strain. Tests are performed at a crosshead speed of 5 mm min^{-1}. We calculate Young's moduli over a range of strains from 0.1% to 0.3%. Single-edge-notch bending (SENB) specimens for fracture toughness experiments are machined from the panels following ASTM standard D5045. Fracture toughness tests are conducted using an Instron 8841 with a 1 kN load cell. Pre-cracks are created by tapping a razor blade into the edge notch of the sample with a hammer. We measure crack lengths optically, using an optical digital microscope (VHX-5000, Keyence). A three-point bending fixture is used with a 12 cm span, and tests are conducted at a crosshead speed of 10 mm min^{-1}. We calculate K_{IC} according to the referred standard.

Fabrication and testing of composite parts. For all composite specimens, we use 12 plies of dry fabric. For FROMP-pDCPD composites, we mix 0.3 molar equivalents of TBP with respect to GC2 with the resin solution. A double-bagged VARTM technique is used to infuse the fabric with the liquid resin. To minimize heat loss during cure and to mitigate quenching of the front, we prepare the layup on a thermally insulating tool plate (448-D, Fibre Glast Developments Corp.). A low vacuum (64.3 kPa) is applied on the inner bag using a vacuum pump (2047B-02, WelchTM DryFastTM) to infuse the resin. A 1.7 kPa vacuum is applied on the outer bag using a second vacuum pump to compact the layup for producing cured parts with a high volume fraction of fibres. To initiate FROMP, we place an electrically resistive wire within the fabric stack and connect it in series to a DC power source. Once resin completely infuses the fabric stack, we apply a constant electric current across the resistive wire until FROMP initiates in the layup, and then turn the power off.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

We evaluate the effect of three different curing modes on total cure time for a 10 cm × 20 cm laminate (Fig. 3a–c). For the in-plane curing modes (Fig. 3a, b), a DC power supply powers the embedded wires for 20 seconds at 19 W per wire. For the through-thickness curing mode (Fig. 3c), an AC power supply (L1010, Staco Energy Products Co.) powers the surface heater (SRFG-408/10, OMEGALUX) for 32 seconds at a maximum rating of 320 W. We use an FLIR SC620 thermal infrared camera to record the surface temperature of the layup and to measure the total cure time for each mode. Given that we carry out thermal imaging on the surface of the FRPC layup, we also embed a thermocouple inside the layup to measure the actual temperature of the FRPC during manufacturing. We found infrared readings to be lower than the actual laminate temperature by around 30 °C, owing to heat dissipation and radiation by the layup materials.

We prepare the corrugated composite panel (Fig. 2i) by stacking 12 plies of 13 cm × 23 cm fabric on a custom metal tool plate. To minimize heat loss to the thermally conductive metal tooling during cure, we preheat the entire layup to 50 °C in a convection oven before resin infusion. The layup is then removed from the oven and immediately infused with resin. Finally, FROMP is triggered using the thermal stimulus via the embedded resistive wire.

We manufacture conventionally cured pDCPD and epoxy composite panels as a control, for comparing the mechanical properties of the FROMP-cured composites. We fabricate these panels using a traditional wet-layup technique and then cure them in a hydraulic press (MTP-13, Tetrahedron) under constant applied platen force of 13.3 kN. The matrix material for the conventional-cured pDCPD-FRPCs is a resin formulation with a pot life of at least one hour (1.0 molar equivalents of TBP with respect to GC2). We use the same cure cycle as for the conventionally cured neat pDCPD panels. The matrix material and cure cycle for the epoxy composites are the same as for the neat epoxy panels described previously.

We determine the void content of the fabricated laminates by polishing the cross-sections of five to eight specimens taken from each FRPC panel, and imaging them using an optical digital microscope (VHX-5000, Keyence). These images are then analysed using ImageJ software to calculate the total void area to cross-sectional area ratios for each specimen. The void volume fraction for each panel is then taken as the average void volume fraction across the analysed specimens. The fibre volume fraction, V_f, of FRPC panels is calculated as $V_f = \frac{f_A n}{\rho_f t}$, where f_A is the areal weight of fabric, n is the number of plies, ρ_f is the fibre density and t is the average laminate thickness. The quality of different FRPC panels is summarized in Extended Data Table 2.

For the tensile tests, we tab carbon FRPC panels with G10 FR4 glass-fibre/epoxy end tabs that are 44.5 mm long, 0.20 mm thick and with a 20° taper. We roughen the surfaces of the end tabs and panels at the adjoining areas by lightly sandblasting them to increase mechanical grip, and adhere them to each other using JB Weld adhesive. The panels are then cut into tensile specimens with nominal dimensions of 203.2 mm × 12.7 mm. We test composite specimens according to ASTM standard D3039 with a crosshead speed of 2 mm min^{-1} on an Instron 5984 with a 150 kN load cell and use an Instron video extensometer to measure the strain. Young's moduli are calculated over a strain range from 0.3% to 0.6%.

Computational modelling. In an axisymmetric setting, the governing partial differential equations describing the FROMP reaction can be written as:

$$\begin{cases} \kappa\left[\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r} + \frac{\partial^2 T}{\partial z^2}\right] + \rho H_r \frac{\partial \alpha}{\partial t} = \rho C_p \frac{\partial T}{\partial t} \\ \frac{\partial \alpha}{\partial t} = A \exp\left(-\frac{E}{RT}\right)(1-\alpha)^n \alpha^m \frac{1}{1+\exp[C(\alpha-\alpha_c)]}, \end{cases}$$

where T (in K) and α (non-dimensional) respectively denote the temperature and the degree of cure; κ (in W m^{-1} K^{-1}) denotes the thermal conductivity; C_p (in J kg^{-1} K^{-1}) is the specific heat; ρ (in kg m^{-3}) denotes the density; H_r (in J kg^{-1}) is the total enthalpy of the reaction; r and z (in m) are the radial and longitudinal coordinates; and t (in s) is the time. The second equation corresponds to the cure kinetics model, with A (8.55×10^{15} s^{-1}), E (110.75 kJ mol^{-1}), R (8.314 J mol^{-1} K^{-1}), n (1.7), and m (0.8) respectively denoting the pre-exponential factor, the activation energy, the universal gas constant, and two constants associated with the Prout–Tompkins model that account for the autoactivating effects. The model also includes an added diffusion factor with $C = 14.5$ and $\alpha_c = 0.4$ to describe the cure kinetics of the monomer along with the effects of diffusion[31]. We calculated the above constants by solving a nonlinear model-fitting optimization problem on the heat of reaction curves obtained by dynamic DSC experiments on a formulation containing 0.5 molar equivalents of TBP with respect to GC2 catalyst. The above governing equations are completed by the following initial and boundary conditions:

$$\begin{aligned} &T(r, z, t=0) = T_0 \\ &\alpha(0 \le r \le a, z, t=0) = \alpha_0 \\ &\{T(0 \le r \le a, z=0, t) = T_{trig} \quad 0 \le t \le t_{trig} \\ &\frac{\partial T}{\partial z}(0 \le r \le a, z=0, t) = 0 \quad t > t_{trig} \end{aligned}$$

while adiabatic conditions are imposed at all other boundaries. T_0 (23 °C) and α_0 (0.07) are respectively the initial temperature and the degree of cure of the monomer solution. T_{trig} is the trigger temperature. The physical and thermal properties used in this study are listed in Extended Data Table 3.

We conducted the numerical analysis using the Multiphysics Object-Oriented Simulation Environment (MOOSE)[32], an open source C++ finite-element solver that includes mesh adaptivity capability. The system of nonlinear partial differential equations is solved at every time step using a combination of the implicit Euler time-stepping scheme and a preconditioned Jacobian-free Newton Krylov method[33,34].

The first simulation presented in Fig. 3d involves the frontal polymerization of DCPD inside a glass tube (Extended Data Fig. 2a). The domain dimensions are length (L) = 7.5 mm, inner radius (a) = 5.5 mm and wall thickness (b) = 1 mm, with L chosen to be long enough to capture the quasi-steady-state propagation of the front. We use 151,470 four-node quadrilateral elements to discretize the domain at the beginning of the simulation, and apply a maximum refinement level of 9 to adapt the mesh in the vicinity of the advancing front. The polymerization is initiated by applying a T_{trig} of 210 °C, for a t_{trig} of 7 seconds, along the left edge of the domain (Extended Data Fig. 2a). Extended Data Fig. 2b, c depict the associated temperature contours during polymerization. The simulation yields a maximum temperature, T_{max}, of 223 °C and a velocity of the polymerization front, V, of 6.4 cm min^{-1}, in good agreement with the experimental results.

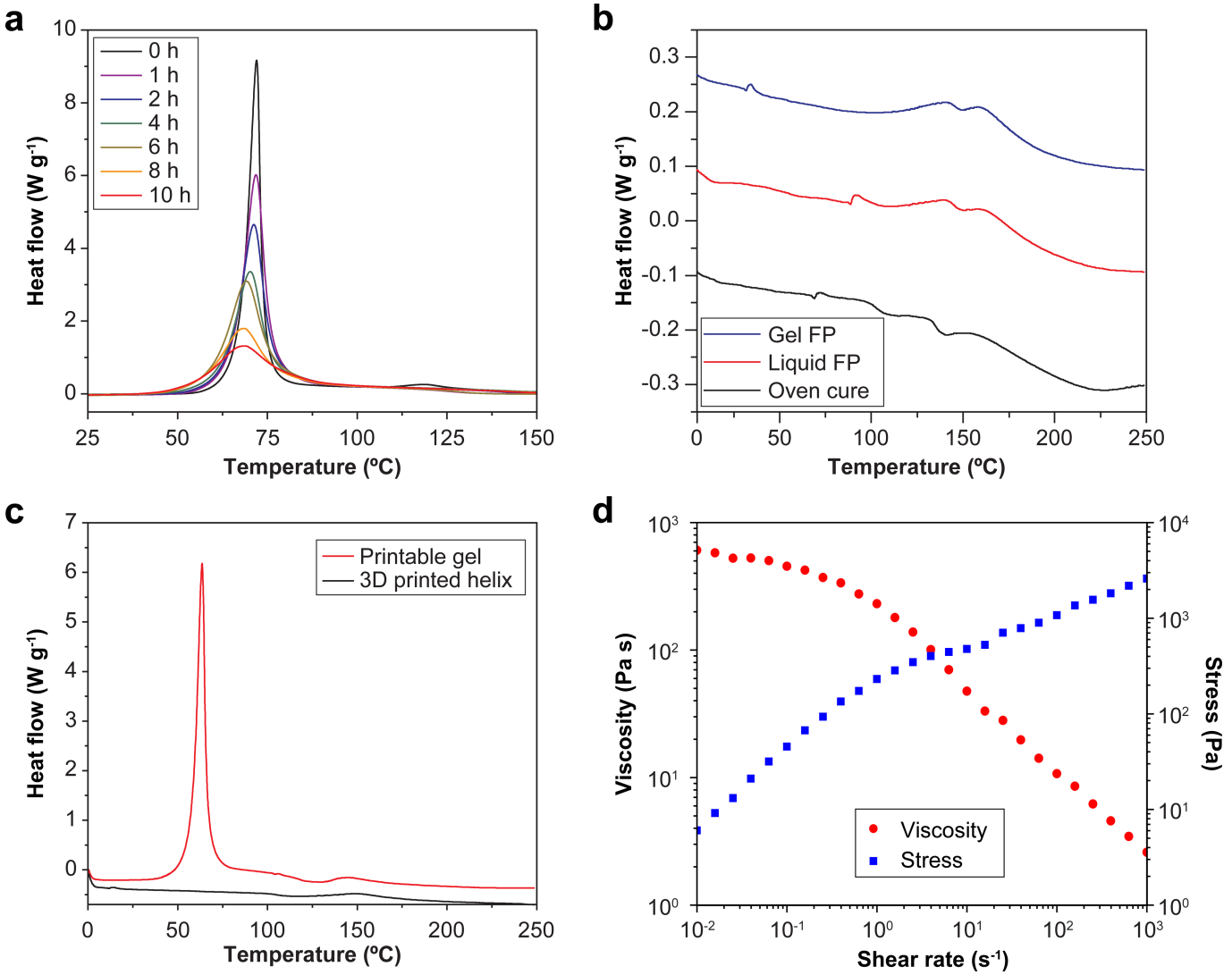
To capture the effect of a carbon-fibre tow as a conductive element on the propagation of the front, we perform a second simulation (Extended Data Fig. 2d), with tow radius (c) = 0.5 mm, and a, b and L the same as for the case without the tow. For this problem, the same mesh is used at the beginning of the simulation; however, a maximum refinement level of 12 is applied to capture the exceptionally sharp gradients associated with the large mismatch in thermal conductivities between the carbon-fibre tow and the DCPD resin ($\kappa_{CF}/\kappa_{resin} = 68.75$). The reaction is initiated by applying a T_{trig} of 210 °C, for a t_{trig} of 7 seconds. Extended Data Fig. 2e, f, which present snapshots of the temperature contours during the frontal-polymerization event, clearly shows the effect of the conductive element on the front profile. The computed front velocity is 10% higher than in the neat resin case.

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

Code availability. We performed the simulations using MOOSE, an open-source C++ finite-element framework developed at Idaho National Laboratory (http://mooseframework.org).

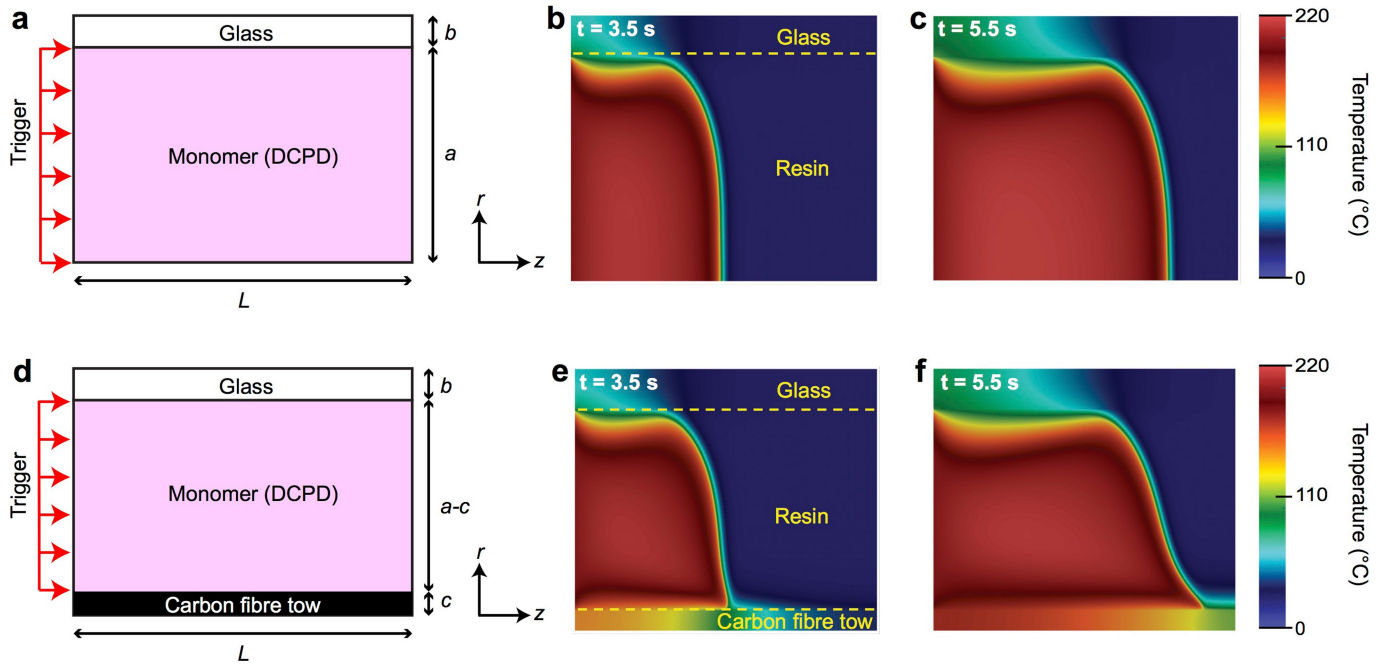
31. Yang, G. & Lee, J. K. Curing kinetics and mechanical properties of endo-dicyclopentadiene synthesized using different Grubbs' catalysts. *Ind. Eng. Chem. Res.* **53**, 3001–3011 (2014).
32. Gaston, D., Newman, C., Hansen, G. & Lebrun-Grandié, D. MOOSE: a parallel computational framework for coupled systems of nonlinear equations. *Nucl. Eng. Des.* **239**, 1768–1778 (2009).
33. Knoll, D. A. & Keyes, D. E. Jacobian-free Newton–Krylov methods: a survey of approaches and applications. *J. Comput. Phys.* **193**, 357–397 (2004).
34. Pernice, M. & Walker, H. NITSOL: a Newton iterative solver for nonlinear systems. *SIAM J. Sci. Comput.* **19**, 302–318 (1998).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 1 | Characterization of DCPD gel and pDCPD products. a, Heat of reaction, measured by DSC, for the formulation presented in Fig. 4b after ageing the resin at 23 °C for the indicated times. As the curing proceeds, the exothermic peak shifts slightly to lower temperatures and broadens. **b**, Heat of reaction of pDCPD specimens produced by liquid FROMP, gel FROMP and conventional (oven) cure approaches, indicating fully cured products with a degree of cure of 99.6%, 99.6% and 99.7%, respectively. **c**, Heat of reaction of the gel before 3D printing and of a cured part after printing, measured by DSC. There is minimal heat of reaction in the printed polymer, indicating a 99.2% degree of cure. **d**, Rheological profile of the 3D printable gel, showing shear thinning behaviour.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 2 | Simulation of FROMP reaction. a, Schematic representation of the axisymmetric model of FROMP in a glass tube. **b**, **c**, Propagation of the reaction front in neat resin. **d**, Schematic representation of the axisymmetric model of FROMP in the presence of a carbon-fibre tow placed at the centre of the glass tube. **e**, **f**, Evolution of the location and profile of the polymerization front in the presence of a carbon-fibre tow.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Inhibitor concentration and resin incubation time for different manufacturing techniques and the corresponding front temperature and velocity

Manufacturing technique	Inhibitor type	Inhibitor concentration (molar equiv. to GC2)	Resin incubation time	Front temperature* (°C)	Front velocity* (cm min^{-1})
Liquid resin (gel time measurements)	TMP, TEP, TBP	0-14	< 1 min	~ 215 (0.3-1 equiv.) ~ 200 (2-4 equiv.) ~ 175 (8 equiv.)†	See Fig. 4a
Stiff gel FROMP	TBP	2	18 h @ room temp.	145	3.6
3D printing	TEP	0.5	160 min @ 23.0 °C	170‡	9.6
FROMP-pDCPD neat panel (liquid FROMP)	TBP	1	< 1 min	216	6.8
FROMP-pDCPD neat panel (gel FROMP)	TBP	1	6 h	184	5.7
Conventional-cured pDCPD neat panel	TBP	1	< 1 min	N/A	N/A
FROMP-pDCPD FRPC	TBP	0.3	< 5 min	138	9.8
Conventional-cured pDCPD FRPC	TBP	1	< 5 min	N/A	N/A

TMP: trimethyl phosphite; TEP: triethyl phosphite; TBP: tributyl phosphite.
*Front temperatures and velocities are highly dependent on the material and geometrical conditions used for each manufacturing technique. The values reported here correspond to the exact experimental conditions described in the Methods section.
†Front temperatures for a given concentration of phosphite inhibitor are similar regardless of the identity of the inhibitor.
‡The front temperature of the 3D printed filament is recorded using a thermal infrared camera, however, it is difficult to determine the exact temperature of the front due to the low distance between the heated substrate and printed material.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Comparison of FRPC panels made with different manufacturing techniques

Manufacturing technique	Void content (%)	V_f (%)	Degree of cure (%)
FROMP-pDCPD FRPC	0.15 ± 0.20	51.3 ± 1.0	80.5 ± 3.9
Conventional-cured pDCPD FRPC	1.30 ± 0.56	51.1 ± 1.3	89.6 ± 1.3
Conventional-cured epoxy FRPC	0.25 ± 0.15	52.2 ± 1.2	98.8 ± 1.7

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 3 | Physical and thermal properties of the various components used in computational modelling

	κ (W m^{-1} K^{-1})	ρ (kg m^{-3})	H_r (J g^{-1})	C_p (J kg^{-1} K^{-1})
Monomer	0.15	980	350	1,600
Carbon fibre	10.45	1,760	_	795
Glass	1.14	2,230	_	800

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0056-8

Nanoscale synthesis and affinity ranking

Nathan J. Gesmundo[1,3], Bérengère Sauvagnat[2], Patrick J. Curran[2], Matthew P. Richards[2], Christine L. Andrews[2], Peter J. Dandliker[2] & Tim Cernak[1,4]*

Most drugs are developed through iterative rounds of chemical synthesis and biochemical testing to optimize the affinity of a particular compound for a protein target of therapeutic interest. This process is challenging because candidate molecules must be selected from a chemical space of more than 10^{60} drug-like possibilities[1], and a single reaction used to synthesize each molecule has more than 10^7 plausible permutations of catalysts, ligands, additives and other parameters[2]. The merger of a method for high-throughput chemical synthesis with a biochemical assay would facilitate the exploration of this enormous search space and streamline the hunt for new drugs and chemical probes. Miniaturized high-throughput chemical synthesis[3–7] has enabled rapid evaluation of reaction space, but so far the merger of such syntheses with bioassays has been achieved with only low-density reaction arrays, which analyse only a handful of analogues prepared under a single reaction condition[8–13]. High-density chemical synthesis approaches that have been coupled to bioassays, including on-bead[14], on-surface[15], on-DNA[16] and mass-encoding technologies[17], greatly reduce material requirements, but they require the covalent linkage of substrates to a potentially reactive support, must be performed under high dilution and must operate in a mixture format. These reaction attributes limit the application of transition-metal catalysts, which are easily poisoned by the many functional groups present in a complex mixture, and of transformations for which the kinetics require a high concentration of reactant. Here we couple high-throughput nanomole-scale synthesis with a label-free affinity-selection mass spectrometry bioassay. Each reaction is performed at a 0.1-molar concentration in a discrete well to enable transition-metal catalysis while consuming less than 0.05 milligrams of substrate per reaction. The affinity-selection mass spectrometry bioassay is then used to rank the affinity of the reaction products to target proteins, removing the need for time-intensive reaction purification. This method enables the primary synthesis and testing steps that are critical to the invention of protein inhibitors to be performed rapidly and with minimal consumption of starting materials.

The merger of nanoscale synthesis with affinity ranking, which we call NanoSAR, could enable data-rich interrogations of reaction–chemical–bioactivity space (Fig. 1). Nanoscale synthesis consumes less than 0.05 mg of substrate per reaction[3], so precious starting materials such as complex drug-like intermediates can be used. Affinity-selection mass spectrometry (ASMS) is a sensitive bioassay[18,19] that relates protein binding to mass detection of a candidate molecule. ASMS was chosen to assay crude nanoscale reaction mixtures directly because misleading readouts related to catalysts or other reaction components can be omitted as long as the molecular masses of the desired products and residual reagents can be resolved by high-resolution mass spectrometry. However, most existing ASMS protocols provide only a binary readout of affinity[20]. We envisioned that affinity could be ranked by titrating down the concentration of protein while maintaining a constant concentration of compound. As the protein concentration decreases, competition for binding is induced[21], and only compounds with high affinity to the target are observed at the lowest protein concentration. With this advance, potent protein inhibitors may be identified at a nanomole-scale while forgoing time-consuming reaction purification.

We validated NanoSAR by synthesizing libraries based on three of the most popular transformations used in pharmaceutical research[22,23]. Each library contains about 20 compounds and each reaction was run on a nanomole scale (less than 0.05 mg). Every analogue was resynthesized on a traditional scale (about 20 mg) so that bioassay results of unpurified nanoscale reactions could be compared directly to data obtained from purified products using a conventional approach. The crude nanoscale reaction mixtures were analysed by using ultra-high-performance liquid chromatography–mass spectrometry (UPLC–MS) to evaluate the reaction efficiency, pooled, and submitted to the ASMS affinity-ranking assay directly (Extended Data Fig. 1). Next, each product was resynthesized on an approximately 20-mg scale using the productive reaction conditions identified on the nanoscale and purified. These purified products were likewise submitted to the ASMS affinity-ranking assay, but were also either submitted to a traditional assay of biochemical function or compared to reported binding affinities. Potent molecules could be identified regardless of the scale of the reaction.

The bioassay results from the crude nanoscale reactions matched those obtained from the purified products. For example, a library of 19 known inhibitors[24] of the kinase ERK2 was prepared by allowing carboxylic acid **1** to react with diverse amines in the presence of HATU (hexafluorophosphate azabenzotriazole tetramethyl uronium) and $^{i}Pr_2NEt$ (19 reactions, Fig. 2a) and submitted to affinity ranking. High-affinity amide product **2** (inhibitory constant $K_i = 35$ nM), which was

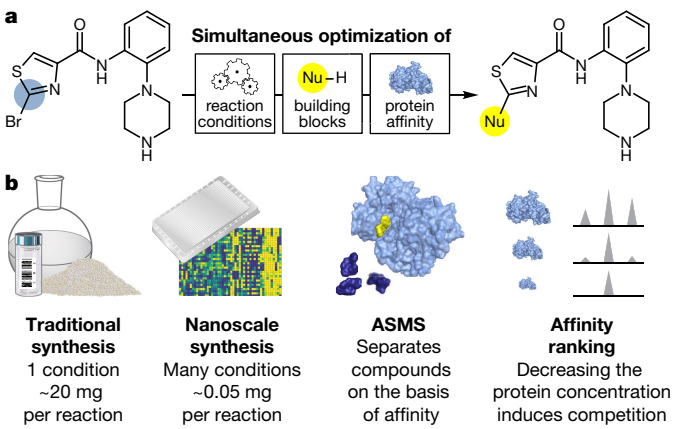


Fig. 1 | Nanoscale synthesis and affinity ranking (NanoSAR). **a**, Reaction space and chemical space are navigated at the nanomole-scale and bioassayed directly to locate potent protein inhibitors and productive reaction conditions simultaneously. **b**, Nanoscale synthesis consumes a fraction of the material that is required for traditional synthesis. Crude reactions are assayed by ASMS, with affinity ranking achieved by decreasing the concentration of the target protein to induce competition among compounds.

[1]Department of Discovery Chemistry, Merck & Co., Inc., Boston, MA, USA. [2]Department of Pharmacology, Merck & Co., Inc., Boston, MA, USA. [3]Present address: Research and Development, AbbVie, North Chicago, IL, USA. [4]Present address: Department of Medicinal Chemistry, Department of Chemistry, Program in Chemical Biology, University of Michigan, Ann Arbor, MI, USA. *e-mail: tcernak@umich.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

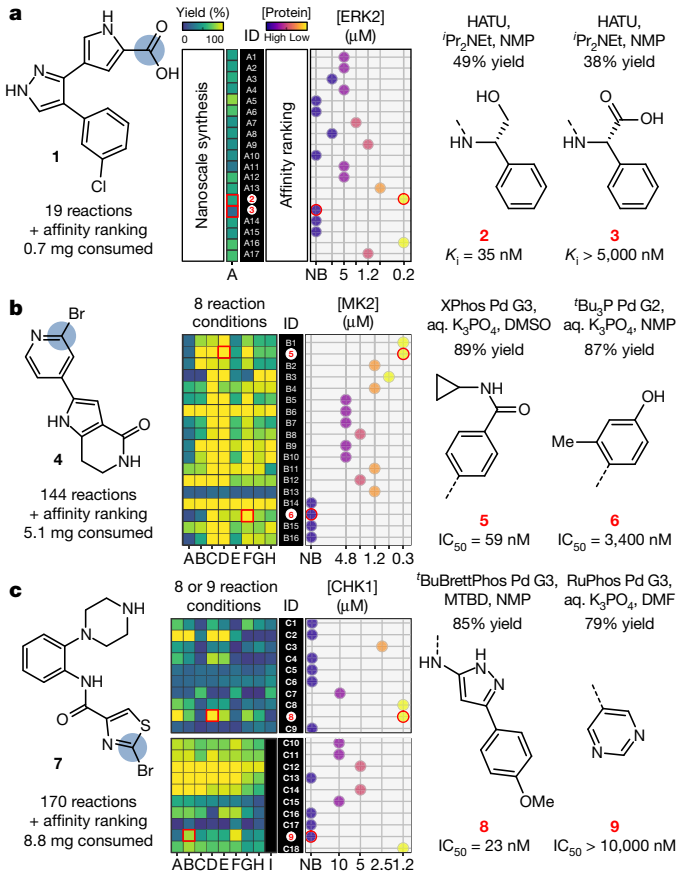


Fig. 2 | NanoSAR identification of ERK2, MK2 and CHK1 inhibitors. Inhibitors were produced using nanoscale synthesis and submitted directly to an ASMS affinity-ranking assay, then resynthesized on a 50-μmol scale, purified, and reassayed by ASMS for comparison to the functional bioactivity (IC_{50}) or reported binding affinity (K_i) of each compound. **a**, A library of 19 ERK2 inhibitors prepared by amide coupling. **b**, A library of 18 MK2 inhibitors prepared by Suzuki coupling. **c**, A library of 20 CHK1 inhibitors prepared by Suzuki and C–N coupling. In each library, every compound has a unique identifier (ID) and was formed under one or more reaction conditions, with the yield, determined by UPLC, for each reaction depicted in the heat map and the affinity-ranking data depicted in the scatter plot. One high-affinity compound (**2**, **5** and **8**) and one low-affinity compound (**3**, **6** and **9**) from each library are shown. The reaction conditions identified on the nanoscale (red box) were then used for the 50-μmol-scale resynthesis of each exemplary compound, and the corresponding affinity-ranking data for that compound are highlighted (red circle). NB indicates no binding was observed in the affinity-ranking assay at the highest protein concentration tested. See Extended Data Figs. 2–4 for reaction conditions (A–I), structures, yields and K_i or IC_{50} values of all compounds.

observed at the lowest ERK2 concentration tested in the affinity-ranking assay, was easily distinguished from low-affinity amide product **3** ($K_i > 5{,}000$ nM), which was not observed at any of the ERK2 concentrations tested. Therefore, with NanoSAR it was possible to differentiate potent from impotent analogues with less than 0.05 mg of material and with no reaction purification (Extended Data Fig. 2).

For this library of ERK2 inhibitors produced by amide coupling, a single reaction condition was sufficient. For subsequent transition-metal-catalysed Suzuki or C–N coupling reactions, we surveyed eight and nine reaction conditions, respectively, to ensure that every product was synthesized successfully. Nanoscale Suzuki coupling of heterocyclic bromide **4** with 18 diverse boronates was achieved with 8 combinations of Buchwald precatalysts[25], bases and solvents (144 reactions, Fig. 2b). Using conditions identified at the nanoscale, all 18 products were prepared on a traditional 50-μmol scale, purified and assayed for inhibition of the function of the kinase MK2[26]. NanoSAR can readily distinguish potent inhibitors, such as **5** (half-maximal inhibitory concentration $IC_{50} = 59$ nM), from impotent inhibitors, such as **6** (IC_{50} 3,400 nM). IC_{50} values for all 18 analogues were determined, using purified compounds produced on a 50-μmol scale, and could be predicted from the NanoSAR data (Extended Data Fig. 3). Similarly, bromide **7** was used to generate 20 distinct inhibitors[27] of the kinase CHK1 via C–C and C–N coupling (170 reactions, Fig. 2c). Potent inhibitors, such as **8** (IC_{50} 23 nM), were readily distinguished from an array of less-potent, lower-affinity compounds, such as **9** ($IC_{50} > 10{,}000$ nM). Once again, affinity-ranking results from crude nanoscale reactions predicted the CHK1 functional activity of purified samples produced on a micromole scale (Extended Data Fig. 4). Samples treated with a metal scavenging resin had low levels of residual palladium; however, simple dilution with water reduced palladium levels to less than 5 p.p.m. (Extended Data Fig. 5).

The information density produced by NanoSAR permits explorations that were previously impractical or impossible, such as the rapid survey of reaction conditions, building blocks and protein affinity (Fig. 3). For example, **7** was coupled to various functionalized nucleophiles via multiple distinct chemical transformations, with affinity information for CHK1 generated for each analogue of the form of **10** produced. As shown in the outer reaction-condition heat map in Fig. 3, we explored combinations of catalysts (**11**–**26**), bases (**27**–**34**) and solvents (*N*-methyl pyrrolidinone (NMP), dimethylsulfoxide (DMSO) and *N*,*N*-dimethylformamide (DMF)) with 17 representative nucleophiles (**D1**–**D17**, Extended Data Fig. 6). For each class of nucleophile, four of the most productive reaction conditions were selected so that **7** could be effectively coupled to 384 diverse building blocks, as shown in the inner heat map of building-block reactivity (Fig. 3). By using four reaction conditions, we identified productive C–N, C–O, C–S or C–C coupling conditions for 345 of 384 products (90%) (Extended Data Figs. 7, 10). The 345 crude reaction mixtures, outlined in black in the inner heat map, were submitted directly to affinity ranking by titrating the CHK1 concentration from 10 μM to 0.2 μM: products that bind tightly to CHK1 appear closer to the centre of the circle in the affinity-ranking plot (Fig. 3). Key compounds were identified for scale-up, purification and analysis in an assay of CHK1 functional activity (Extended Data Fig. 8). Exemplars (**35**–**41**) are shown in Fig. 3 (see Extended Data Fig. 9 for dose titration curves), with each class of functional group revealing unique structure–reactivity and structure–bioactivity relationships. Compounds **38**, **39** and **41** were isolated in 98%, 52% and 24% yield, respectively, from reactions run on a 50-μmol scale, and were potent CHK1 inhibitors, as predicted by NanoSAR.

Subtle reactivity trends were illuminated in the data analysis. For instance, one of the thiol nucleophiles that we tested in the initial survey of reaction conditions (3-phenylpropane-1-thiol, **D15**) couples to **7** in high yield regardless of the catalyst, base or solvent used (see Extended Data Fig. 6 for reaction-condition mapping). This finding suggests that alkyl thiols couple to **7** via a nucleophilic aromatic substitution (S_NAr) mechanism rather than via palladium catalysis. Guided by this information, two of the four reaction conditions selected to couple thiols to **7** used no catalyst. Compound **35** was isolated in 98% yield on a 50-μmol scale using no catalyst, with P_2Et in DMF; however, it was inactive against CHK1, as predicted by NanoSAR.

Higher-order trends also emerge from the data. For example, comparing the reaction performance of boronate to amine building blocks suggests that, at least with **7**, the Suzuki coupling is a more robust reaction than is the C–N coupling. Years of anecdotal experience with these two transformations led us to hypothesize this disparity in reaction robustness[7], and here the phenomenon can be readily visualized. Patterns of bioactivity are also apparent in the central plot of affinity ranking in Fig. 3. For instance, many of the boronate-, amide- and amine-derived products were bioactive, exhibiting affinity to CHK1 and appearing as points closer to the centre of the circle; these compounds exhibited functional inhibition of CHK1 function

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

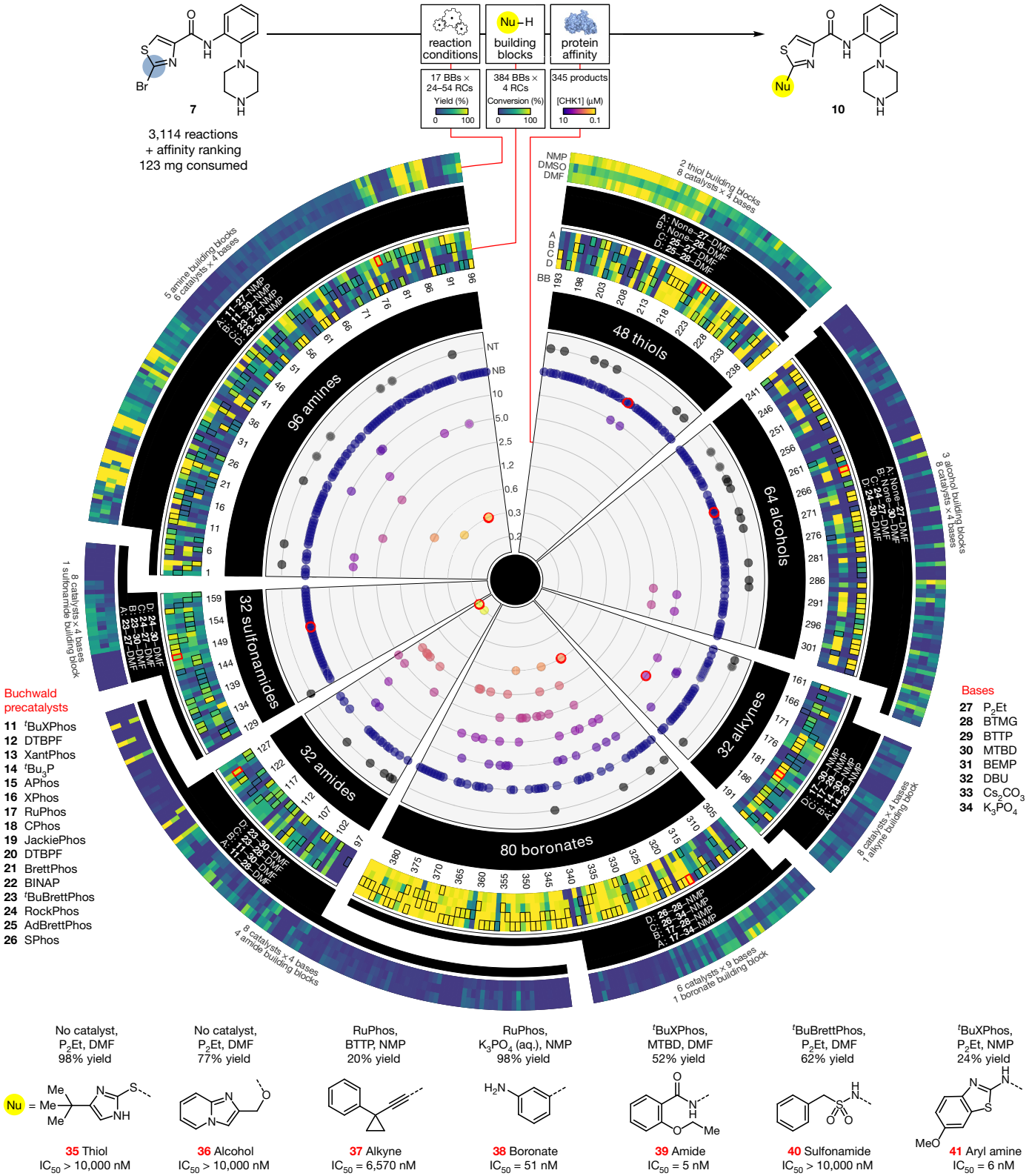


Fig. 3 | NanoSAR mapping of reactivity and bioactivity for diverse CHK1 inhibitors. 3,114 reactions were run, using only 123 mg of **7**, to map structure–reactivity and structure–bioactivity relationships. The outer heat map shows a survey of reaction conditions (RCs) with 17 representative building blocks (BB; **D1**–**D17**, Extended Data Fig. 6). The inner heat map shows library synthesis of 384 diverse building blocks (see Supplementary Information for structures) using four productive reaction conditions (A–D) selected from the reaction-condition survey. Using the reaction conditions outlined in black in the inner heat map, 345 compounds were submitted directly to the ASMS affinity-ranking assay against CHK1, as shown in the central plot. Points closer to the centre of the circle represent compounds whose ASMS binding signals persisted at lower CHK1 concentrations, corresponding to higher CHK1 affinity. NT, not tested owing to 0% yield. NB, no binding observed at 10 μM CHK1. Exemplary compounds (**35**–**41**), the NanoSAR data for which are outlined in red, were resynthesized on a 50-μmol scale and purified to obtain IC_{50} values from a functional assay of CHK1 activity. See Extended Data Figs. 6–9 for details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

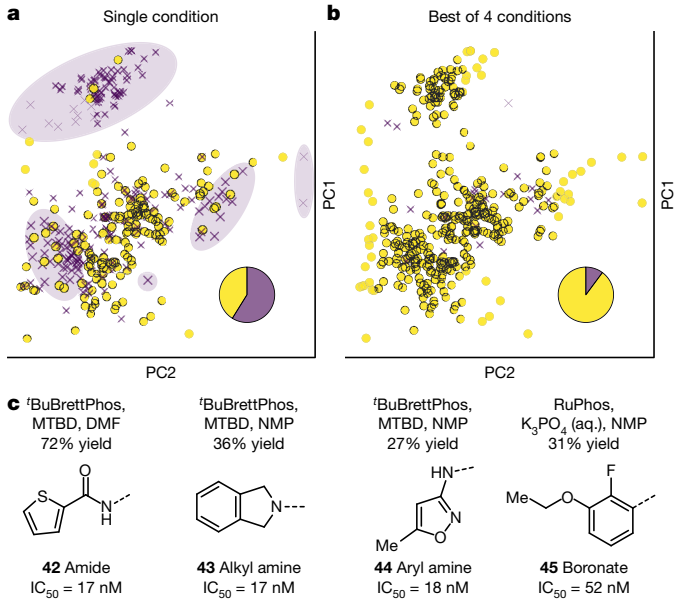


Fig. 4 | NanoSAR enhances coverage of chemical space. a, When coupling **7** to diverse nucleophiles (**D18**–**D402**), a single reaction condition (**11**, **27**, NMP, room temperature) produced 158 (yellow circles) out of 384 (41%; see inset pie chart) designed products, with 226 failures (0% yield; purple crosses). The purple highlighted regions indicate 'dark' regions of principal component (PC) chemical space, where reactions failed. **b**, Matching each nucleophile to the best of four reaction conditions produced 345 out of 384 (90%) products. **e**, 21 compounds from the 'dark' space that were made accessible by NanoSAR bound to CHK1, such as **42**–**45**. CHK1 IC_{50} values are from a functional assay of purified compounds.

(Extended Data Fig. 9). By contrast, few of the sulfonamide-, thiol-, alkyne- or alcohol-derived products bound to CHK1, and these linkages could reasonably be deprioritized from future design cycles. In total, only 123 mg of **7** were consumed to run 3,114 reactions, to match 345 of the 384 diverse nucleophiles with coupling conditions, to assay the products and to paint a landscape of reactivity and bioactivity.

For comparison, coupling of bromide **7** to the same 384 diverse nucleophiles under a single reaction condition—the recently reported combination of ^{t}BuXPhos Pd G3 (**11**)[25] with P_2Et (**27**) as a base[28], which is among the most substrate-tolerant palladium-catalysed coupling conditions known[7]—yielded detectable product (**10**) for only 158 of 384 reactions (41%). Therefore, screening reaction conditions by using nanoscale synthesis enabled deeper exploration of chemical space, with an additional 187 targeted compounds synthesized and assayed successfully (Fig. 4). Under a single reaction condition, synthesis was unsuccessful for large swaths of chemical space and for compounds on the fringes of the space that we studied (Fig. 4a). NanoSAR identified 21 compounds from this 'dark' space that bound to CHK1 (Extended Data Fig. 10). Exemplary compounds were resynthesized at a 50-μmol scale, purified and tested in a functional assay, whereupon several products (such as **42**–**45**) displayed functional inhibition of CHK1 at low, nanomolar concentrations (Fig. 4c). As can be seen from Fig. 3, a large number of reactions produce no observable product and a large number of compounds have no affinity to the protein target, which, in our experience, is typical of drug discovery. Machine-learning algorithms may one day facilitate the navigation of these problems[29,30]; however, locating bioactive molecules and conditions for their synthesis for now remains an experimental science, which NanoSAR is well positioned to accelerate.

Data availability

The data that support the findings of this study are available within the paper and its Supplementary Information. Source Data for Figs. 2–4 and Extended Data Figs. 2–6, 8–10 are available with the online version of the paper.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0056-8.

Received: 2 October 2017; Accepted: 14 February 2018;
Published online 23 April 2018.

1. Reymond, J. L. The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).
2. Murray, P. M., Tyler, S. N. G. & Moseley, J. D. Beyond the numbers: charting chemical reaction space. *Org. Process Res. Dev.* **17**, 40–46 (2013).
3. Buitrago Santanilla, A. et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015).
4. Shevlin, M. Practical high-throughput experimentation for chemists. *ACS Med. Chem. Lett.* **8**, 601–607 (2017).
5. Collins, K. D., Gensch, T. & Glorius, F. Contemporary screening approaches to reaction discovery and development. *Nat. Chem.* **6**, 859–871 (2014).
6. Troshin, K. & Hartwig, J. F. Snap deconvolution: an informatics approach to high-throughput discovery of catalytic reactions. *Science* **357**, 175–181 (2017).
7. Kutchukian, P. S. et al. Chemistry informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chem. Sci.* **7**, 2604–2613 (2016).
8. Werner, M. et al. Seamless integration of dose-response screening and flow chemistry: efficient generation of structure–activity relationship data of β-secretase (BACE1) inhibitors. *Angew. Chem. Int. Ed.* **53**, 1704–1708 (2014).
9. Desai, B. et al. Rapid discovery of a novel series of Abl kinase inhibitors by application of an integrated microfluidic synthesis and screening platform. *J. Med. Chem.* **56**, 3033–3047 (2013).
10. Guetzoyan, L., Nikbin, N., Baxendale, I. R. & Ley, S. V. Flow chemistry synthesis of zolpidem, alpidem and other GABAA agonists and their biological evaluation through the use of in-line frontal affinity chromatography. *Chem. Sci.* **4**, 764–769 (2013).
11. Karageorgis, G., Dow, M., Aimon, A., Warriner, S. & Nelson, A. Activity-directed synthesis with intermolecular reactions: development of a fragment into a range of androgen receptor agonists. *Angew. Chem. Int. Ed.* **54**, 13538–13544 (2015).
12. Murray, J. B., Roughley, S. D., Matassova, N. & Brough, P. A. Off-rate screening (ORS) by surface plasmon resonance. An efficient method to kinetically sample hit to lead chemical space from unpurified reaction products. *J. Med. Chem.* **57**, 2845–2850 (2014).
13. Baranczak, A. et al. Integrated platform for expedited synthesis–purification–testing of small molecule libraries. *ACS Med. Chem. Lett.* **8**, 461–465 (2017).
14. Price, A. K., MacConnell, A. B. & Paegel, B. M. hνSABR: photochemical dose–response bead screening in droplets. *Anal. Chem.* **88**, 2904–2911 (2016).
15. Vastl, J., Wang, T., Trinh, T. B. & Spiegel, D. A. Encoded silicon-chip-based platform for combinatorial synthesis and screening. *ACS Comb. Sci.* **19**, 255–261 (2017).
16. Goodnow, R. A. Jr, Dumelin, C. E. & Keefe, A. D. DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nat. Rev. Drug Discov.* **16**, 131–147 (2017).
17. Annis, D. A. et al. An affinity selection–mass spectrometry method for the identification of small molecule ligands from self-encoded combinatorial libraries. Discovery of a novel antagonist of *E. coli* dihydrofolate reductase. *Int. J. Mass Spectrom.* **238**, 77–83 (2004).
18. Andrews, C. L., Ziebell, M. R., Nickbarg, E. & Yang, X. in *Protein and Peptide Mass Spectrometry in Drug Discovery* (eds Gross, M. L. et al.) 253–286 (John Wiley & Sons, Hoboken, 2012).
19. O'Connell, T. N., Ramsay, J., Rieth, S. F., Shapiro, M. J. & Stroh, J. G. Solution-based indirect affinity selection mass spectrometry—a general tool for high-throughput screening of pharmaceutical compound libraries. *Anal. Chem.* **86**, 7413–7420 (2014).
20. Annis, D. A. et al. A general technique to rank protein-ligand binding affinities and determine allosteric versus direct binding site competition in compound mixtures. *J. Am. Chem. Soc.* **126**, 15495–15503 (2004).
21. Cuozzo, J. W. et al. Discovery of a potent BTK inhibitor with a novel binding mode by using parallel selections with a DNA-encoded chemical library. *ChemBioChem* **18**, 864–871 (2017).
22. Schneider, M. et al. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *J. Med. Chem.* **59**, 4385–4402 (2016).
23. Brown, D. G. & Boström, J. Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? *J. Med. Chem.* **59**, 4443–4458 (2016).
24. Aronov, A. M. et al. Flipped out: structure-guided design of selective pyrazolylpyrrole ERK inhibitors. *J. Med. Chem.* **50**, 1280–1287 (2007).
25. Bruno, N. C., Tudge, M. T. & Buchwald, S. L. Design and preparation of new palladium precatalysts for C–C and C–N cross-coupling reactions. *Chem. Sci.* **4**, 916–920 (2013).
26. Anderson, D. R. et al. Pyrrolopyridine inhibitors of mitogen-activated protein kinase-activated protein kinase 2 (MK-2). *J. Med. Chem.* **50**, 2647–2654 (2007).
27. Huang, X. et al. Structure-based design and optimization of 2-aminothiazole-4-carboxamide as a new class of CHK1 inhibitors. *Bioorg. Med. Chem. Lett.* **23**, 2590–2594 (2013).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

28. Buitrago Santanilla, A. et al. P_2Et phosphazene: a mild, functional group tolerant base for soluble, room temperature Pd-catalyzed C–N, C–O, and C–C cross-coupling reactions. *Org. Lett.* **17**, 3370–3373 (2015).
29. Schneider, P. & Schneider, G. De novo design at the edge of chaos. *J. Med. Chem.* **59**, 4077–4086 (2016).
30. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross coupling using machine learning. *Science* **360**, 186–190 (2018).

Acknowledgements We are grateful to L. Nogle, D. Smith and M. Pietrafitta (MSD) for assistance with compound purification, S. Bano and X. Bu (MSD) for measuring residual palladium concentrations, A. M. Norris (MSD) for assistance with graphics and Z. Gu (DFKZ German Cancer Research Center) for suggestions for data visualization. N.J.G. was supported by an MRL Postdoctoral Research Fellowship.

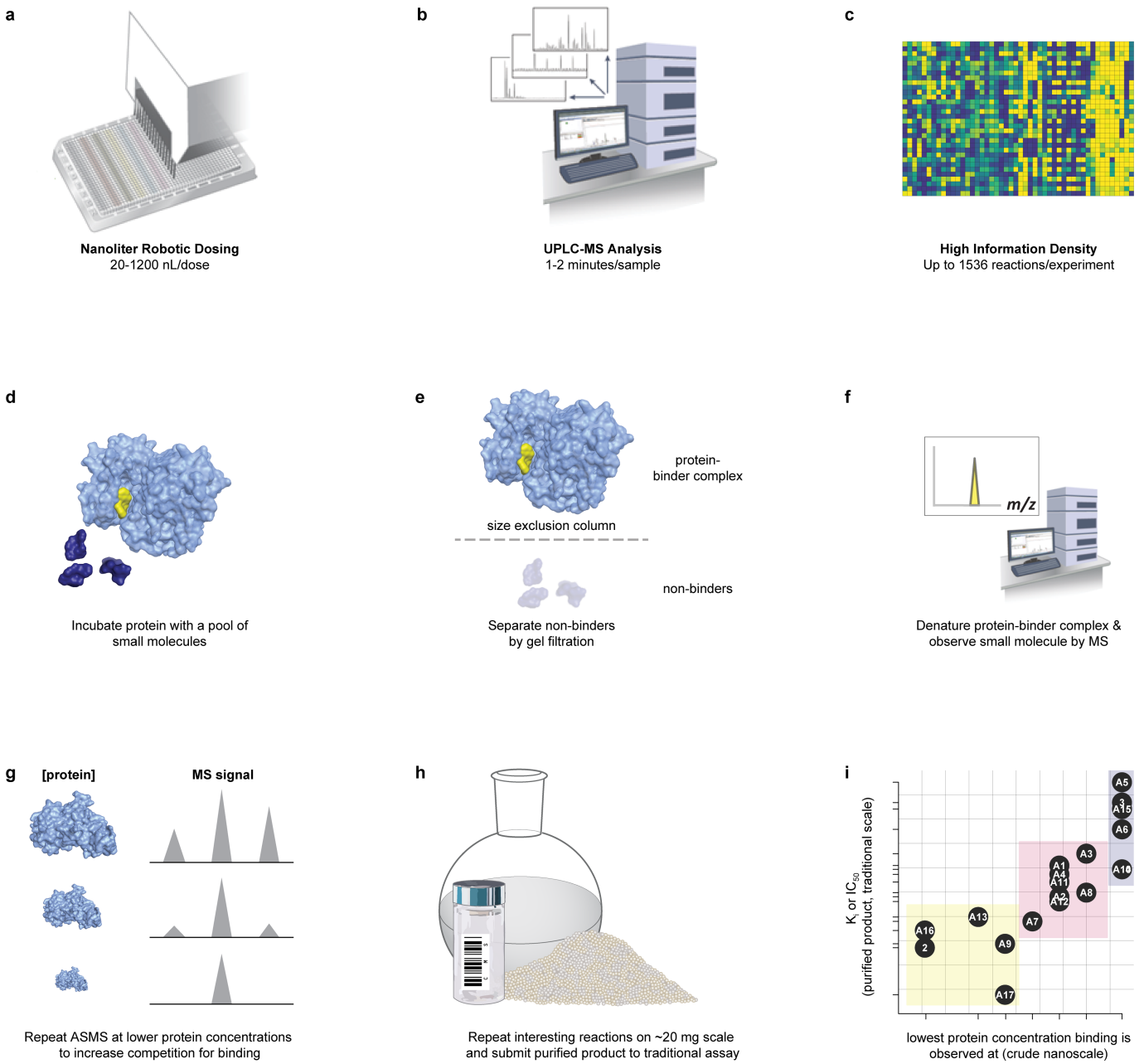
Reviewer information *Nature* thanks J. Janey, D. Young and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions N.J.G., P.J.D. and T.C. conceived the study and wrote the manuscript. N.J.G., B.S., P.J.C., P.J.D. and T. C. designed and executed the experiments and analysed the data. C.L.A. and M.P.R. developed the protein titration method. T.C. supervised the work.

Competing interests The authors declare no competing interests.

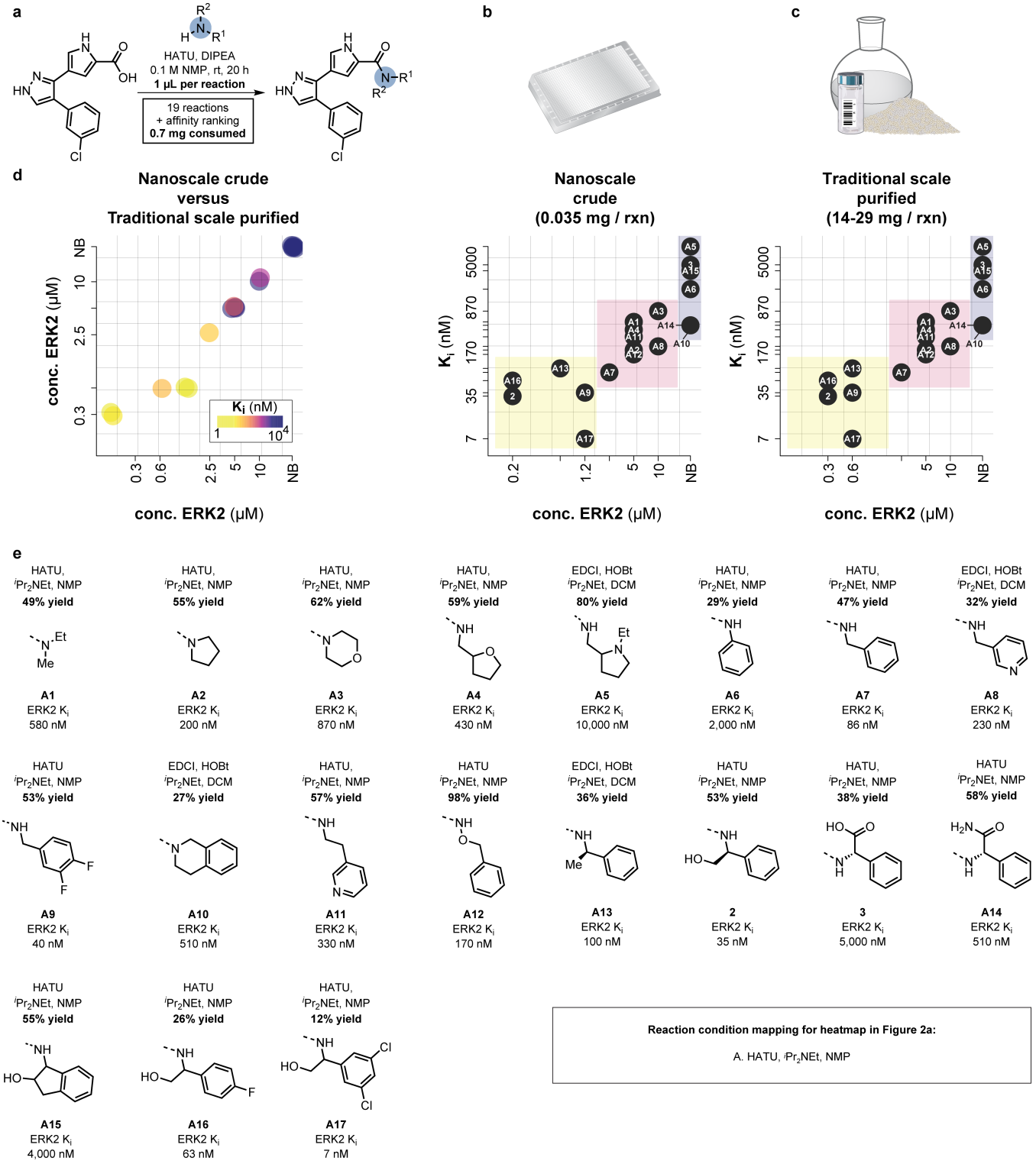
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0056-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0056-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



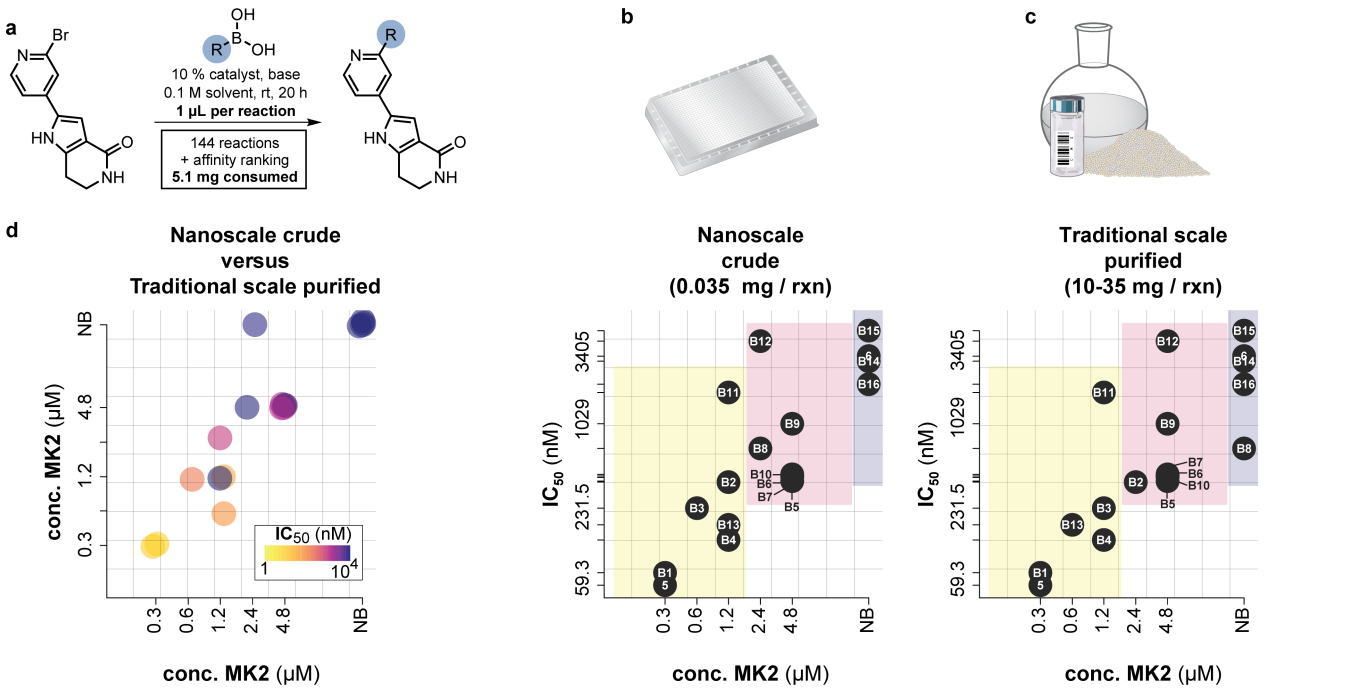
Extended Data Fig. 1 | Additional details of the nanoscale synthesis and affinity ranking workflow. a, Reaction solutions are dosed using a TTP mosquito liquid-handling robot that handles increments of 20 nl. **b**, **c**, UPLC–MS analysis confirms the presence of products from crude reactions (**b**), which is visualized in heat maps (**c**). **d**, For the ASMS assay, productive reactions are mass-encoded, pooled and incubated with a protein of interest. **e**, **f**, Elution through a size-exclusion column separates protein-bound compounds from unbound compounds (**e**), and binders can be observed by high-performance liquid chromatography–mass spectrometry (HPLC–MS) following denaturation of the protein–ligand complex (**f**). **g**, Decreasing the concentration of protein in the ASMS assay increases the competition for binding so that ligands can be categorized as high, medium or low affinity. **h**, Scale-up of compounds and purification by using a traditional approach. **i**, Measurement of IC_{50} in a functional assay enables comparison of the new method to existing assay technology. See Supplementary Information for additional details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 2 | Structures, yields, K_i values and ASMS comparison plots for the ERK2 library. a, A library of 19 unique ERK2 inhibitors (**2**, **3**, **A1**–**A17**) was produced by coupling **1** to diverse amines under a single reaction condition (HATU, ^{i}Pr$_2$NEt, NMP; see Fig. 2a). **b**, Crude nanoscale reactions were submitted to the ASMS assay, with affinity ranking determined by reducing the concentration of ERK2 to increase the competition for binding. The lowest ERK2 concentration at which the mass signal of the compound was still observed is shown on the abscissa. The K_i value reported[24] from purified product samples generated on a 500-times-larger reaction scale is shown on the ordinate. The colour shading groups compounds into low affinity (purple), modest affinity (magenta) or high affinity (yellow), as determined by the ASMS assay. rxn, reaction. **c**, Comparison of the results of the ASMS affinity-ranking assay (abscissa) to the reported ERK2 K_i values (ordinate). Both datasets are from purified product samples produced on a 50-µmol scale. **d**, Comparison of results of ASMS affinity ranking from crude reaction mixtures generated on a 100-nmol scale with those from purified product samples generated on a 50-µmol scale. Points are coloured by K_i, and jittering was applied to this categorical data to reveal overlapping data. **e**, Product structures with reaction conditions used, isolated yields and K_i values. See Supplementary Information for additional details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



e

B1 — XPhos Pd G3, aq. K_3PO_4, DMSO — **71% yield** — MK2 IC_{50} 74 nM

5 — XPhos Pd G3, aq. K_3PO_4, DMSO — **89% yield** — MK2 IC_{50} 59 nM

B2 — XPhos Pd G3, aq. K_3PO_4, DMSO — **12% yield** — MK2 IC_{50} 367 nM

B3 — XPhos Pd G3, aq. K_3PO_4, DMSO — **25% yield** — MK2 IC_{50} 232 nM

B4 — XPhos Pd G3, aq. K_3PO_4, DMSO — **74% yield** — MK2 IC_{50} 132 nM

B5 — XPhos Pd G3, aq. K_3PO_4, DMSO — **79% yield** — MK2 IC_{50} 374 nM

B6 — XPhos Pd G3, aq. K_3PO_4, DMSO — **100% yield** — MK2 IC_{50} 399 nM

B7 — XPhos Pd G3, aq. K_3PO_4, DMSO — **72% yield** — MK2 IC_{50} 410 nM

B8 — XPhos Pd G3, aq. K_3PO_4, DMSO — **91% yield** — MK2 IC_{50} 665 nM

B9 — XPhos Pd G3, aq. K_3PO_4, DMSO — **57% yield** — MK2 IC_{50} 1,029 nM

B10 — XPhos Pd G3, aq. K_3PO_4, DMSO — **15% yield** — MK2 IC_{50} 421 nM

B11 — XPhos Pd G3, aq. K_3PO_4, DMSO — **31% yield** — MK2 IC_{50} 1,790 nM

B12 — XPhos Pd G3, aq. K_3PO_4, DMSO — **70% yield** — MK2 IC_{50} 4,435 nM

B13 — XPhos Pd G3, aq. K_3PO_4, DMSO — **68% yield** — MK2 IC_{50} 172 nM

B14 — XPhos Pd G3, aq. K_3PO_4, DMSO — **34% yield** — MK2 IC_{50} 3,110 nM

6 — ^{t}Bu3P Pd G2, aq. K_3PO_4, NMP — **87% yield** — MK2 IC_{50} 3,405 nM

B15 — XPhos Pd G3, aq. K_3PO_4, DMSO — **71% yield** — MK2 IC_{50} 5,340 nM

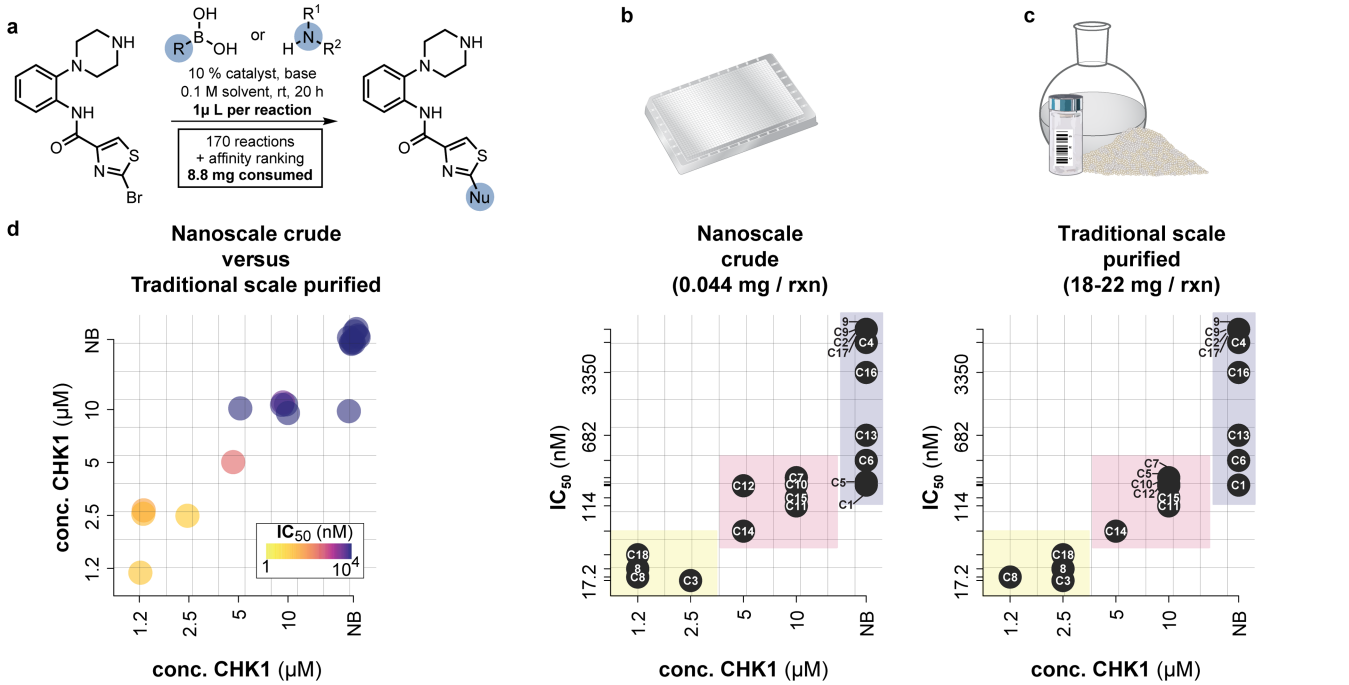
B16 — XPhos Pd G3, aq. K_3PO_4, DMSO — **55% yield** — MK2 IC_{50} 2,070 nM

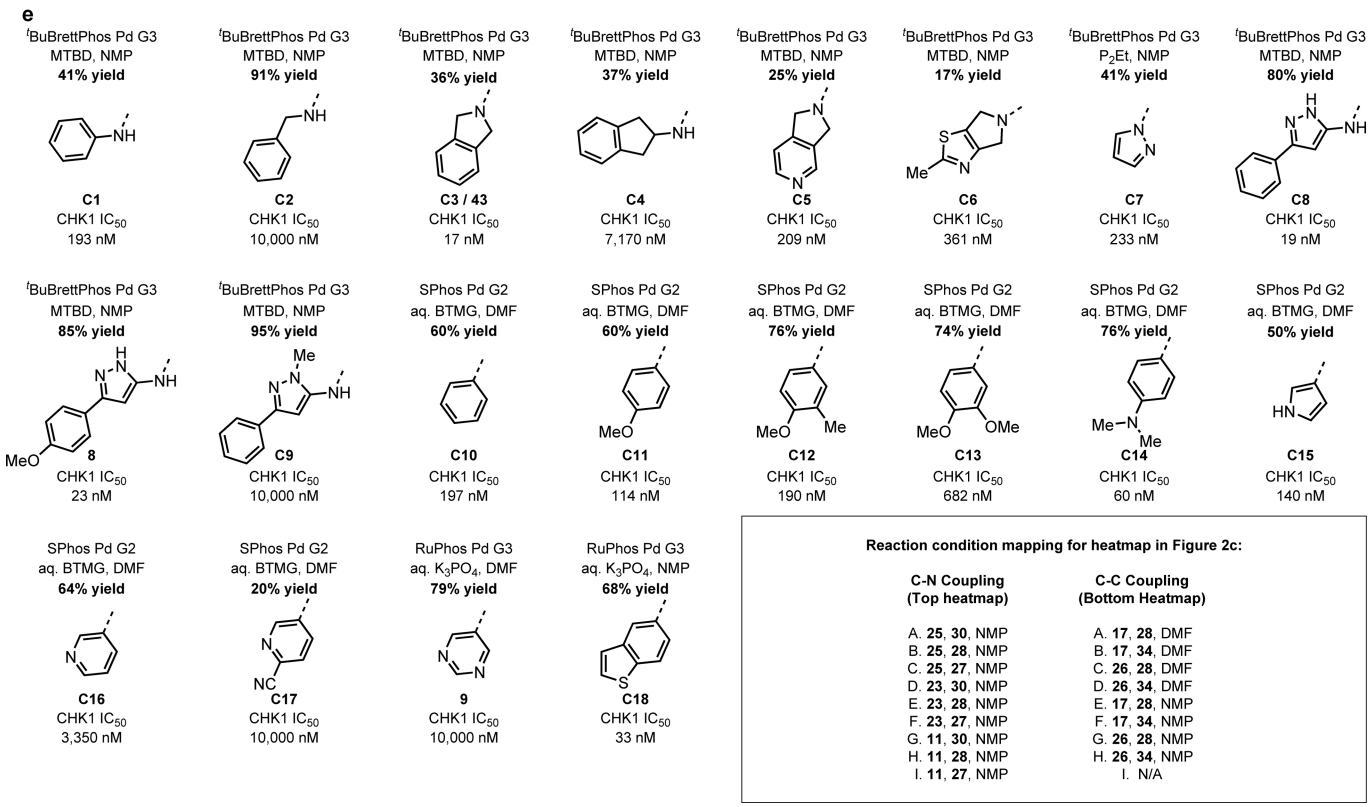
Reaction condition mapping for heatmap in Figure 2b:

A. **14**, **28**, DMSO	E. **14**, **28**, NMP
B. **14**, **34**, DMSO	F. **14**, **34**, NMP
C. **16**, **28**, DMSO	G. **16**, **28**, NMP
D. **16**, **34**, DMSO	H. **16**, **34**, NMP

Extended Data Fig. 3 | Structures, yields, IC_{50} values and ASMS comparison plots for the MK2 Library. a, A library of 18 unique MK2 inhibitors (**5**, **6**, **B1**–**B16**) was produced by coupling **4** to diverse boronates under eight reaction conditions (see Fig. 2b). **b**–**e**, As in Extended Data Fig. 2, but for MK2 and IC_{50} values instead of ERK2 and K_i values.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



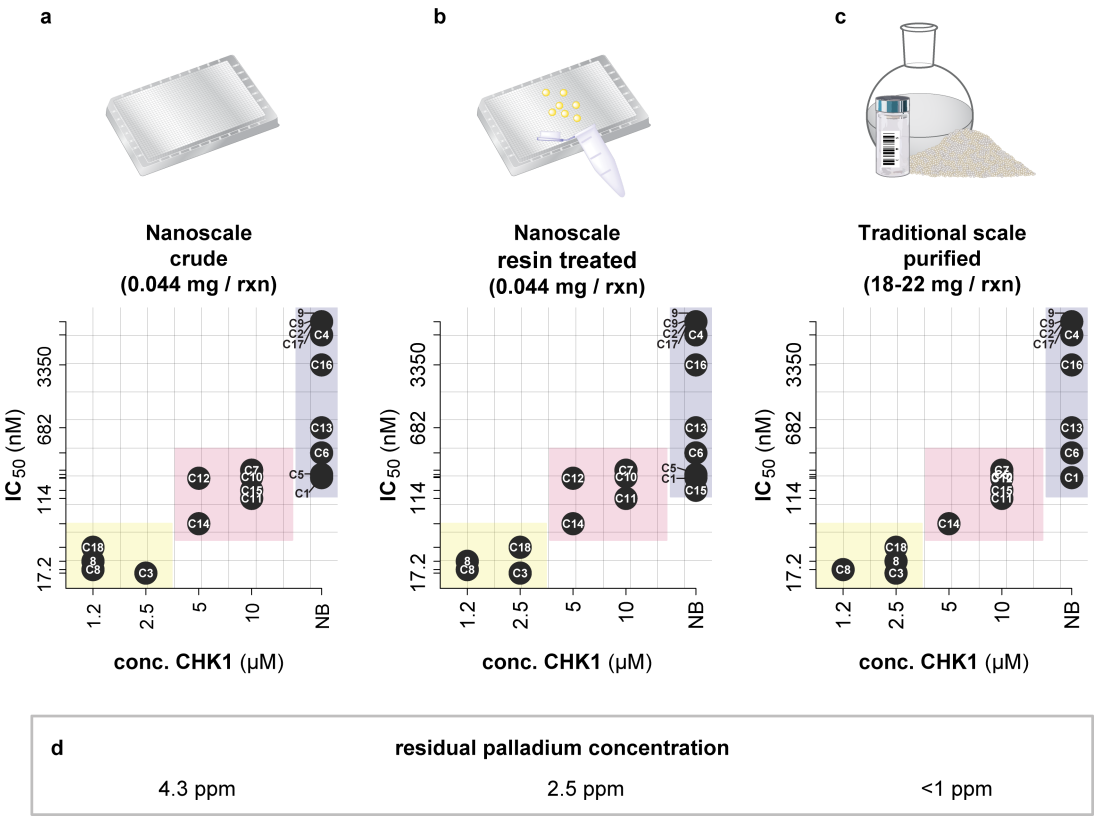


Reaction condition mapping for heatmap in Figure 2c:

C-N Coupling (Top heatmap)	C-C Coupling (Bottom Heatmap)
A. **25**, **30**, NMP	A. **17**, **28**, DMF
B. **25**, **28**, NMP	B. **17**, **34**, DMF
C. **25**, **27**, NMP	C. **26**, **28**, DMF
D. **23**, **30**, NMP	D. **26**, **34**, DMF
E. **23**, **28**, NMP	E. **17**, **28**, NMP
F. **23**, **27**, NMP	F. **17**, **34**, NMP
G. **11**, **30**, NMP	G. **26**, **28**, NMP
H. **11**, **28**, NMP	H. **26**, **34**, NMP
I. **11**, **27**, NMP	I. N/A

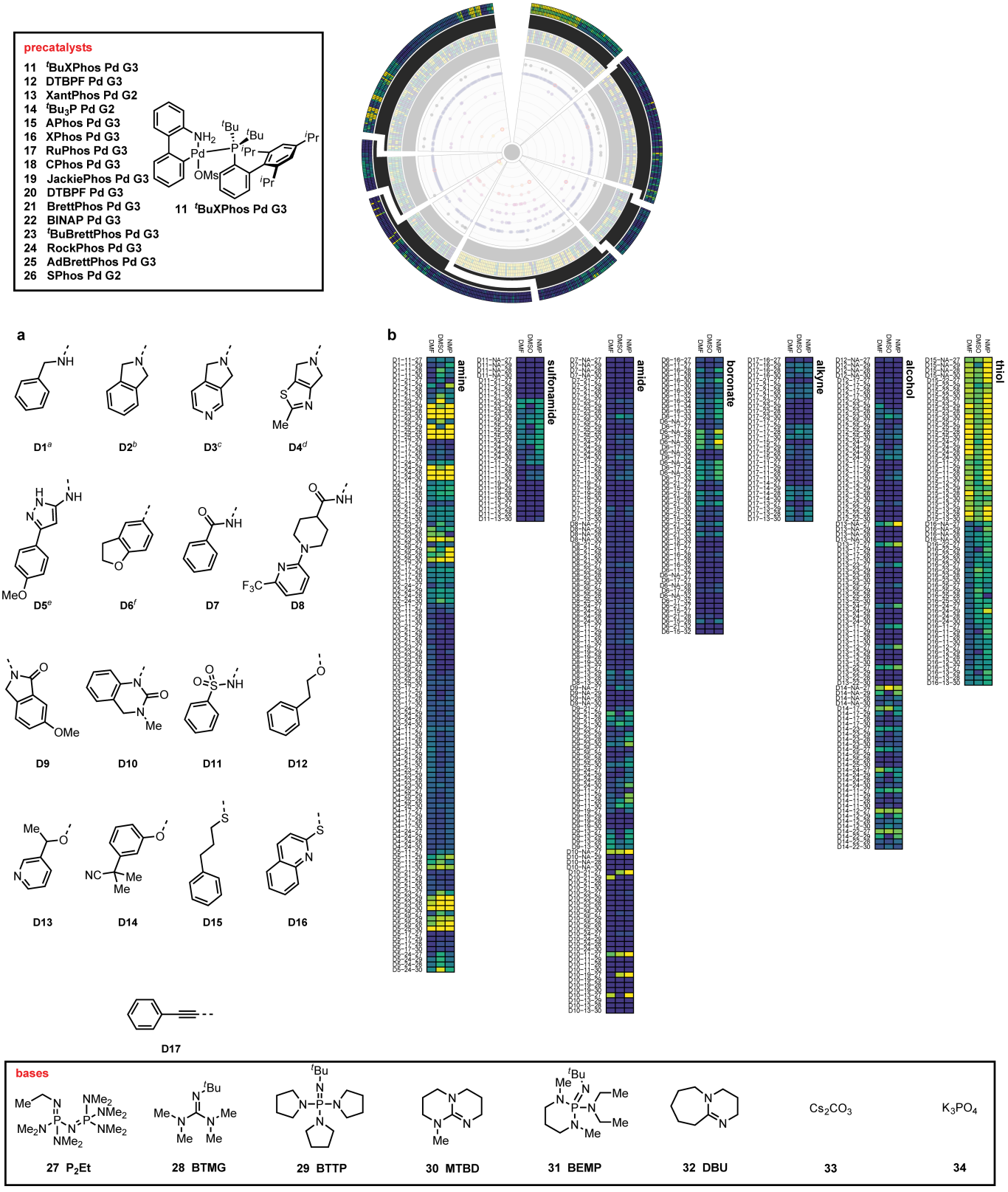
Extended Data Fig. 4 | Structures, yields, IC_{50} values and ASMS comparison plots for the CHK1 Library. a, A library of 20 unique CHK1 inhibitors (**8**, **9**, **C1**–**C18**) was produced by coupling **7** to diverse boronates and amines under eight or nine reaction conditions (see Fig. 2c). **b**–**e**, As in Extended Data Fig. 2, but for CHK1 and IC_{50} values instead of ERK2 and K_i values.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



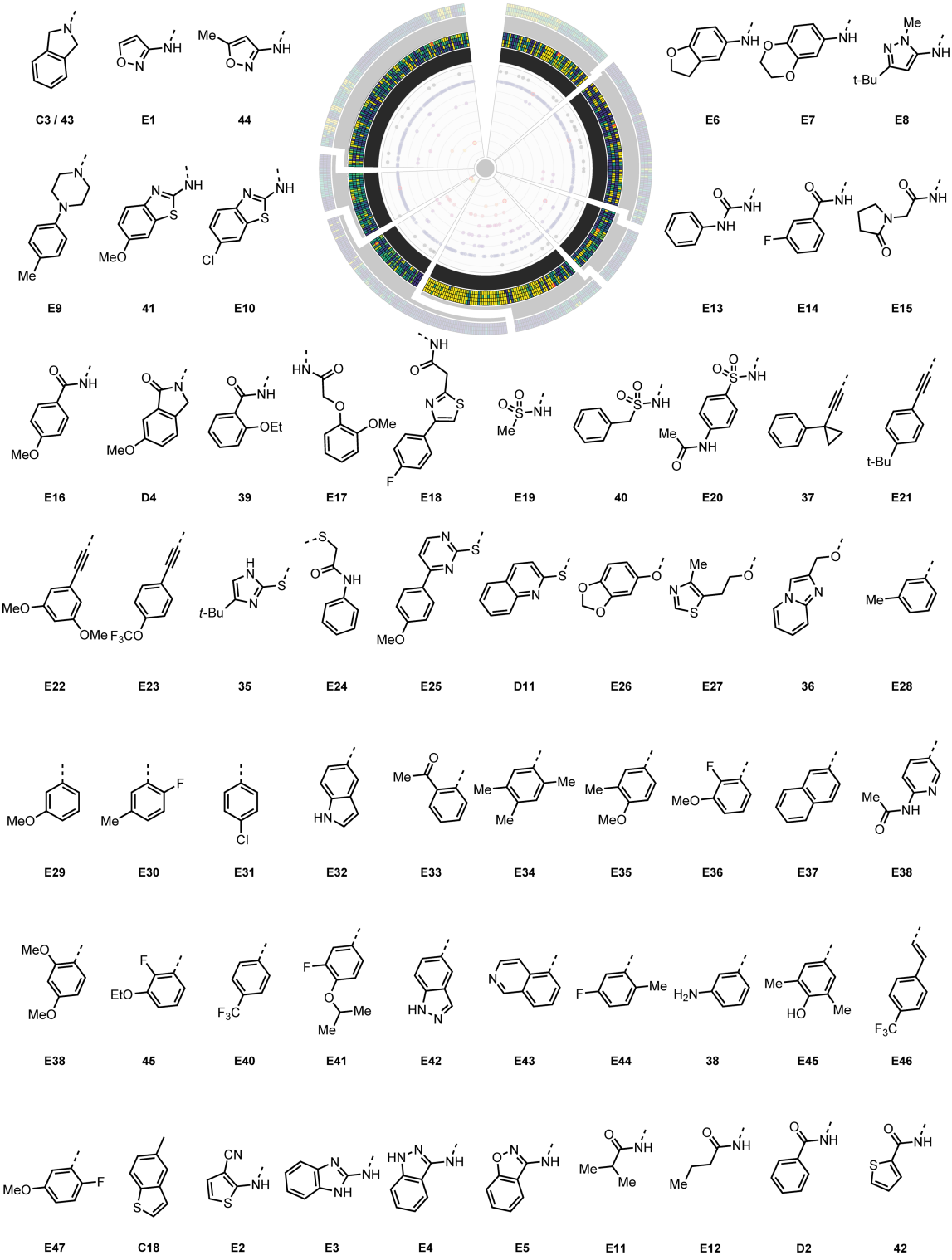
Extended Data Fig. 5 | ASMS comparison plots for different reaction workup protocols and residual levels of palladium from the CHK1 library. Reactions with different workup protocols were submitted to the ASMS assay, with affinity ranking determined by reducing the concentration of CHK1 to increase the competition for binding. The axes of the plots are as in Extended Data Fig. 4b, c. The reaction and workup protocols are as follows: **a**, nanoscale reactions submitted to the affinity-ranking assay with no purification; **b**, nanoscale reactions submitted to the affinity-ranking assay following treatment with SiliaMetS dimercaptotriazine (DMT) resin; and **c**, 50-mmol-scale reactions in which the product samples were purified by reverse-phase HPLC before submission to the affinity-ranking assay. See Supplementary Information for additional details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



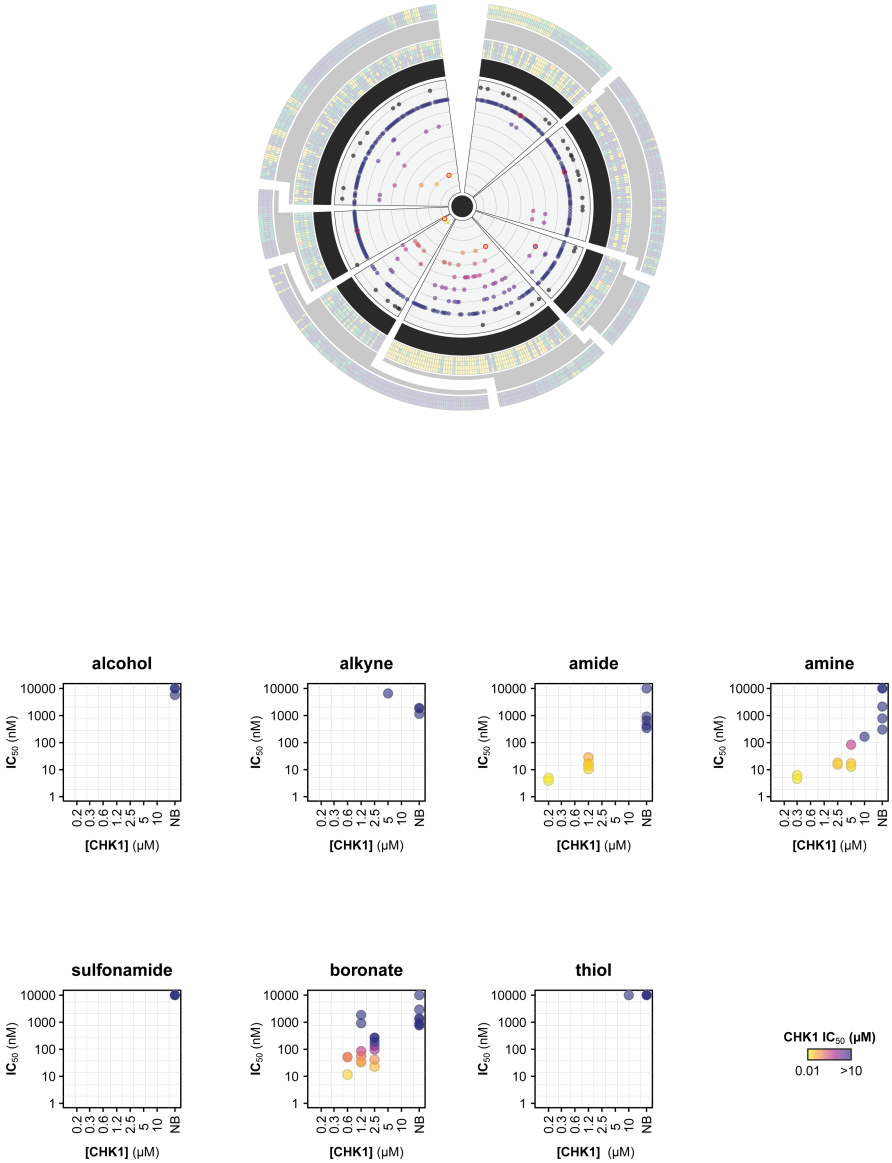
Extended Data Fig. 6 | Catalyst–base survey. See Fig. 3. **a**, Diverse model nucleophiles (**D1**–**D17**) were screened against combinations of catalysts (**11**–**26**) and bases (**27**–**34**) in NMP, DMSO or DMF solvent. **b**, The details beside the heat maps show the mapping of nucleophiles, solvents, catalysts and bases. NA, no catalyst used. Reactions were run on a 100-nmol scale and analysed by UPLC–MS for conversion to products of the form of **10** compared to a biphenyl internal standard. Productive reaction conditions from this screen were selected for use in the subsequent library synthesis campaign. [a]**D1** is compound **C2** in Fig. 2c; [b]**D2** is compound **C3** in Fig. 2c and **43** in Fig. 4; [c]**D3** is compound **C5** in Fig. 2c; [d]**D4** is compound **C6** in Fig. 2c; [e]**D5** is compound **8** in Fig. 2c; [f]prepared from the boronic acid. See Supplementary Information for additional details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



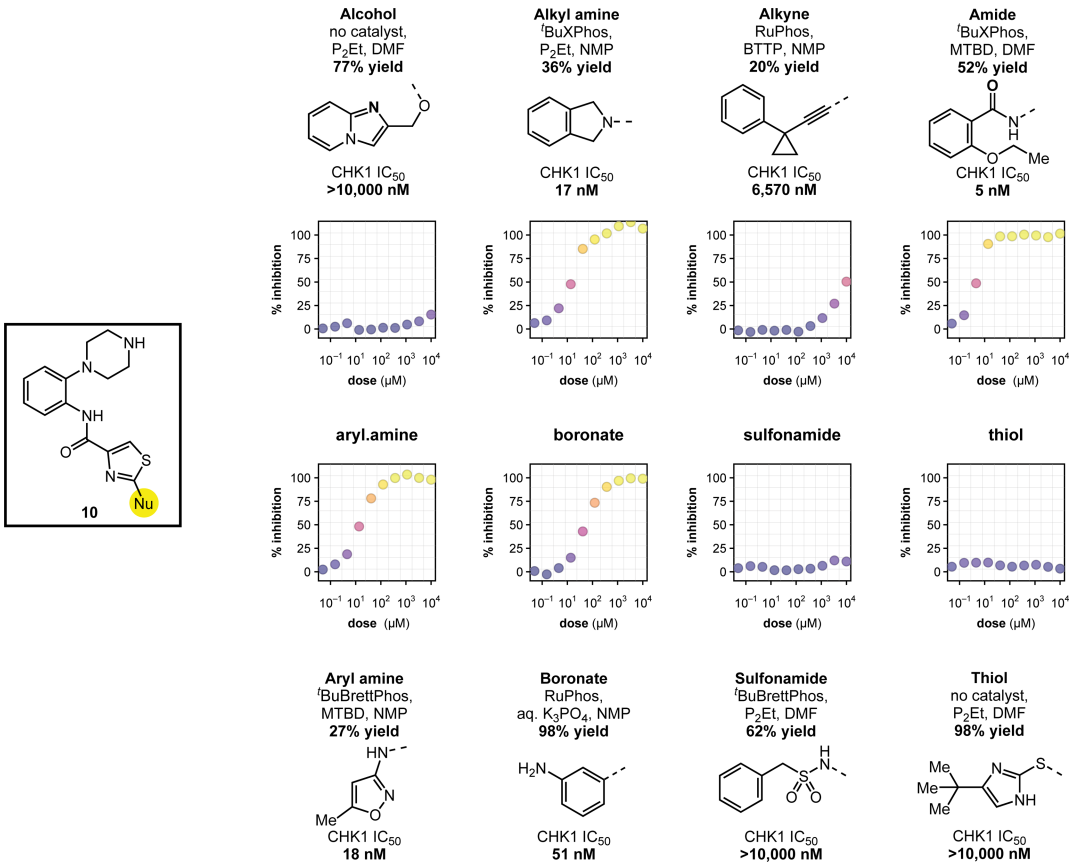
Extended Data Fig. 7 | Exemplary compounds from the synthesis of the library. See Fig. 3. Structures for 62 diverse coupling products (of the form of **10**) are shown, which were selected from the 345 productive reactions identified in a library synthesis campaign targeting 384 products (Fig. 3). Diverse nucleophiles were coupled to **7** using the four reaction conditions selected in Fig. 3 and Extended Data Fig. 6. See Supplementary Information for additional details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



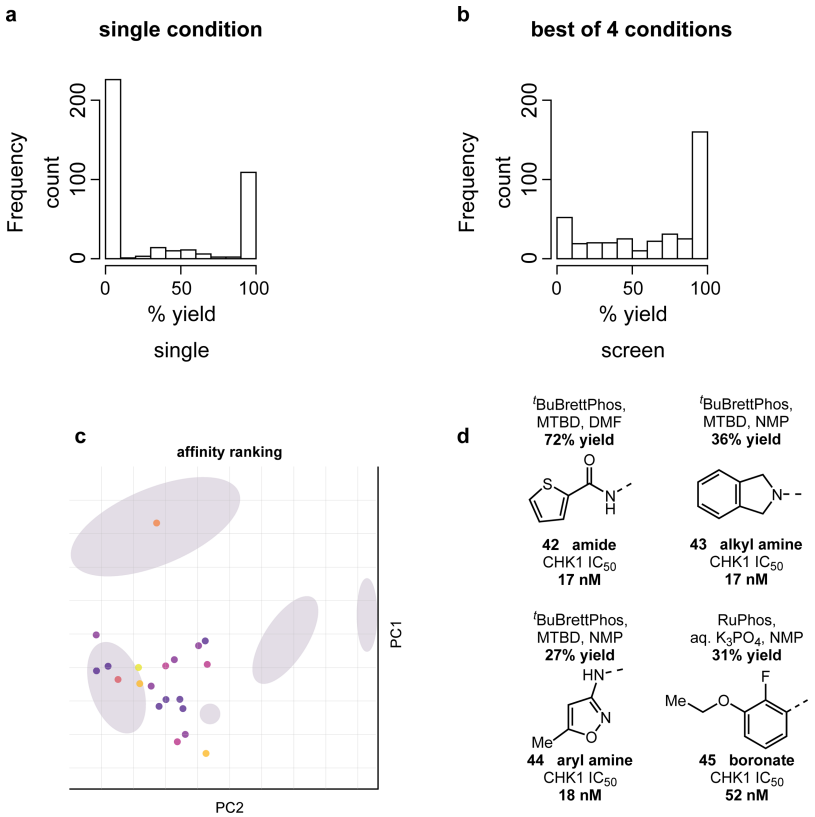
Extended Data Fig. 8 | Comparison of the results of affinity ranking with IC_{50} values for inhibition of CHK1 functional activity. The 62 exemplary compounds (Extended Data Fig. 6) were selected from 345 that were submitted to affinity ranking (Fig. 3). Crude nanoscale reactions were submitted to the ASMS assay, with affinity ranking determined by reducing the concentration of CHK1 to increase the competition for binding. The axes of the plots are as in Extended Data Fig. 4b, c. Points are coloured by the IC_{50} value for inhibition of CHK1 function (as in Extended Data Fig. 4d). See Supplementary Information for additional details.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Dose response curves for exemplary compounds. The compounds shown are **35**–**41** and **43**. The inhibition of CHK1 functional activity by diverse coupling products (**10**) from the reaction of **7** with nucleophiles under various reaction conditions is shown. Isolated yields were achieved when the reaction conditions shown were used. The dose response curves and IC_{50} values shown were measured on purified product samples generated on a 50-μmol scale and could be predicted from affinity-ranking results of crude nanoscale reactions as shown in Extended Data Fig. 8. See Supplementary Information for dose response curves of additional compounds.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Reaction metrics and chemical 'dark' space. **a**, Histogram of reaction performance for 384 diverse coupling reactions with bromide **7** using a single reaction condition (10 mol% ^{t}BuXPhos Pd G3, P2Et, NMP). Only 158 of the 384 targeted products were observed by mass spectrometry and the majority of the reactions failed (0% conversion to product). **b**, Histogram of reaction performance for 384 diverse coupling reactions with bromide **7** using the best of four reaction conditions, as described in Extended Data Figs. 6 and 7. Of the 384 targeted products, 345 were observed by mass spectrometry, with a more even distribution of reaction yields and the majority of reactions succeeding (100% conversion to product). **c**, Principal-component (PC) analysis of chemical 'dark' space, with each dot representing a compound that was not formed under a single reaction condition (0% yield) but that had affinity to CHK1. By using the best of four reaction conditions, 187 additional products were produced and assayed that bound to CHK1; the colour of the dots reflects the affinity ranking of the compounds. The boundaries of this space are identical to those depicted in Fig. 4a, in which purple shading highlights additional regions where the majority of reactions failed (0% yield). Areas where dots are shown in purple shaded regions depict products with affinity to CHK1 that were formed in 0% yield in **a** but in more than 1% yield in **b**. **d**, Potent CHK1 inhibitors that were produced in 0% yield under a single condition (10 mol% ^{t}BuXPhos Pd G3, P2Et, NMP), but in modest to good yields following reaction-condition screening.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0072-8

Earliest known hominin activity in the Philippines by 709 thousand years ago

T. Ingicco[1,2,3,4]*, G. D. van den Bergh[5], C. Jago-on[6], J.-J. Bahain[1,2,3,4], M. G. Chacón[1,2,3,4,7,8], N. Amano[9], H. Forestier[1,2,3,4], C. King[6], K. Manalo[10], S. Nomade[11,12,13], A. Pereira[1,2,3,4,11,12,13,14,15], M. C. Reyes[6,10]*, A.-M. Sémah[1,2,3,4,16], Q. Shao[17], P. Voinchet[1,2,3,4], C. Falguères[1,2,3,4], P. C. H. Albers[18], M. Lising[6,19], G. Lyras[20], D. Yurnaldi[21], P. Rochette[22,23,24,25,26], A. Bautista[6] & J. de Vos[18]

Over 60 years ago, stone tools and remains of megafauna were discovered on the Southeast Asian islands of Flores, Sulawesi and Luzon, and a Middle Pleistocene colonization by *Homo erectus* was initially proposed to have occurred on these islands[1–4]. However, until the discovery of *Homo floresiensis* in 2003, claims of the presence of archaic hominins on Wallacean islands were hypothetical owing to the absence of in situ fossils and/or stone artefacts that were excavated from well-documented stratigraphic contexts, or because secure numerical dating methods of these sites were lacking. As a consequence, these claims were generally treated with scepticism[5]. Here we describe the results of recent excavations at Kalinga in the Cagayan Valley of northern Luzon in the Philippines that have yielded 57 stone tools associated with an almost-complete disarticulated skeleton of *Rhinoceros philippinensis*, which shows clear signs of butchery, together with other fossil fauna remains attributed to stegodon, Philippine brown deer, freshwater turtle and monitor lizard. All finds originate from a clay-rich bone bed that was dated to between 777 and 631 thousand years ago using electron-spin resonance methods that were applied to tooth enamel and fluvial quartz. This evidence pushes back the proven period of colonization[6] of the Philippines by hundreds of thousands of years, and furthermore suggests that early overseas dispersal in Island South East Asia by premodern hominins took place several times during the Early and Middle Pleistocene stages[1–4]. The Philippines therefore may have had a central role in southward movements into Wallacea, not only of Pleistocene megafauna[7], but also of archaic hominins.

The most recent recoveries in Flores[8,9] and Sulawesi[10] (Indonesia) provide a unique documentation of overseas hominin dispersal during the early Middle Pleistocene epoch. An early presence in the Philippine archipelago has been hypothesized since the 1950s, with the reporting of presumably Pleistocene megafaunal remains and 'Palaeolithic' industries consisting of chopping tools and flakes (the 'Cabalwanian' and 'Liwanian' industries, respectively) from surface finds and excavations in the Cagayan Valley basin of northern Luzon[3,4]. Despite the fact that these early discoveries took place more than 60 years ago, no direct association between megafauna and lithic industries has been documented since then, and no secure numerical dating of both fossil fauna and lithics has been available for this region[11]. To date, the discovery of a human metatarsal in Callao Cave in northern Luzon[6], directly dated to 66.7 ± 1.0 thousand years ago (ka), represented the oldest evidence of the peopling of the Philippines.

In 2013, a survey of the Cagayan Valley near the Rizal Municipality (Kalinga Province) led to the discovery of a concentration of vertebrate bones and stone artefacts scattered on the surface near what became our new excavation site. The Kalinga site (17° 33′ 45.0318″ N, 121° 33′ 35.7372″ E) (Fig. 1b) has been excavated annually since 2014 and has resulted in the discovery of in situ megafauna and associated stone artefacts. The substrate consists of the upper part of the Awidon Mesa Formation, a 400-m thick sequence of alluvial stream deposits (mainly sandstones and claystones) intercalated with volcaniclastic and pyroclastic layers (Fig. 1a). These sediments were deposited on an alluvial fan system in braided streams of the paleo-Chico River as a consequence of uplift in the Central Cordillera to the west[12,13]. During a poorly constrained Pleistocene phase of folding in response to east–west compression, alluvial fan deposition in the Kalinga area came to a halt.

We conducted the main 16-m^2 excavation at the head of a modern, dry stream valley, north of a small hill and down to a maximum depth of 2 m (Fig. 1c and Supplementary Information). A 25×1-m^2 slot trench was excavated down the hill to the main excavation. Together, these excavations revealed a total of 7.5 m of stratigraphy comprising four main sedimentary units, in ascending order: unit A, unit F, unit G and unit J (Fig. 1d, e and Extended Data Fig. 1). An almost-complete disarticulated skeleton of *R. philippinensis* (Extended Data Fig. 2) was found embedded in the basal sediments of unit F lying across the base of an erosional channel surface that cuts down vertically into sandy unit A. This channel was filled with an up to 3.25-m thick mudflow (unit F; see Extended Data Fig. 3, 4), which covered the bones, along with an in situ tektite as well as 57 stone tools and sparse fossils of other animals (Geoemydidae, *Varanus* cf. *salvator*, *Stegodon* cf. *luzonensis* and *Cervus* cf. *mariannus*) (see Supplementary Information). The archaeological layer (unit F) is conformably overlaid by an approximately 1.15-m thick, sterile, cross-bedded coarse sandy fluvial unit with silty lenses (unit G), which is in turn conformably overlaid by unit H, a 2.5-m thick silty pedogenized layer with rhyzoliths.

The 57 stone artefacts account for six cores, 49 flakes and two possible hammer stones that all originated from unit F (Fig. 2 and Supplementary Information). With the exception of the two possible hammer stones (Fig. 2b), all artefacts lack a patinated lustre and have a fresh appearance, indicating that any transport was minimal. The knapping strategies were oriented towards short and unorganized core reduction, resulting in non-standardized flake morphologies and dimensions, and all artefacts lacked any intentional retouch.

[1]UMR 7194, CNRS, Paris, France. [2]Département Homme et Environnement, Muséum national d'Histoire naturelle, Paris, France. [3]Université de Perpignan Via Domitia, Perpignan, France. [4]Sorbonne Université, Paris, France. [5]Centre for Archaeological Science, School of Earth and Environmental Sciences, University of Wollongong, Wollongong, New South Wales, Australia. [6]National Museum of the Philippines, Manila, Philippines. [7]IPHES – Institut Català de Paleoecologia Humana i Evolució Social, Tarragona, Spain. [8]Area de Prehistoria, Universitat Rovira i Virgili (URV), Tarragona, Spain. [9]Max Planck Institute for the Science of Human History, Jena, Germany. [10]Archaeological Studies Program, University of the Philippines Diliman, Quezon City, Philippines. [11]Laboratoire des Sciences du Climat et de l'Environnement, CEA, Gif Sur Yvette Cedex, France. [12]UMR 8212, CNRS, Gif Sur Yvette Cedex, France. [13]Université Paris-Saclay, Gif Sur Yvette Cedex, France. [14]Ecole française de Rome, Roma, Italy. [15]Sezione di scienze preistoriche e antropologiche, Dipartimento di Studi Umanistici, Università degli Studi di Ferrara, Ferrara, Italy. [16]Institut de Recherche pour le Développement, UMR LOCEAN 7159, Bondy, France. [17]School of Geographical Sciences, Nanjing Normal University, Nanjing, China. [18]Naturalis Biodiversity Center, Leiden, Netherlands. [19]Department of Sociology and Anthropology, Ateneo de Manila University, Quezon City, Philippines. [20]Faculty of Geology and Geoenvironment National and Kapodistrian University of Athens, Athens, Greece. [21]Centre for Geological Survey, Geological Agency, Bandung, Indonesia. [22]Aix-Marseille Université, Aix en Provence, France. [23]UM34 CNRS, Aix en Provence, France. [24]IRD, Aix en Provence, France. [25]Collège de France, Paris, France. [26]CEREGE, Aix en Provence, France. *e-mail: ingicco@mnhn.fr; mariancreyes@gmail.com

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

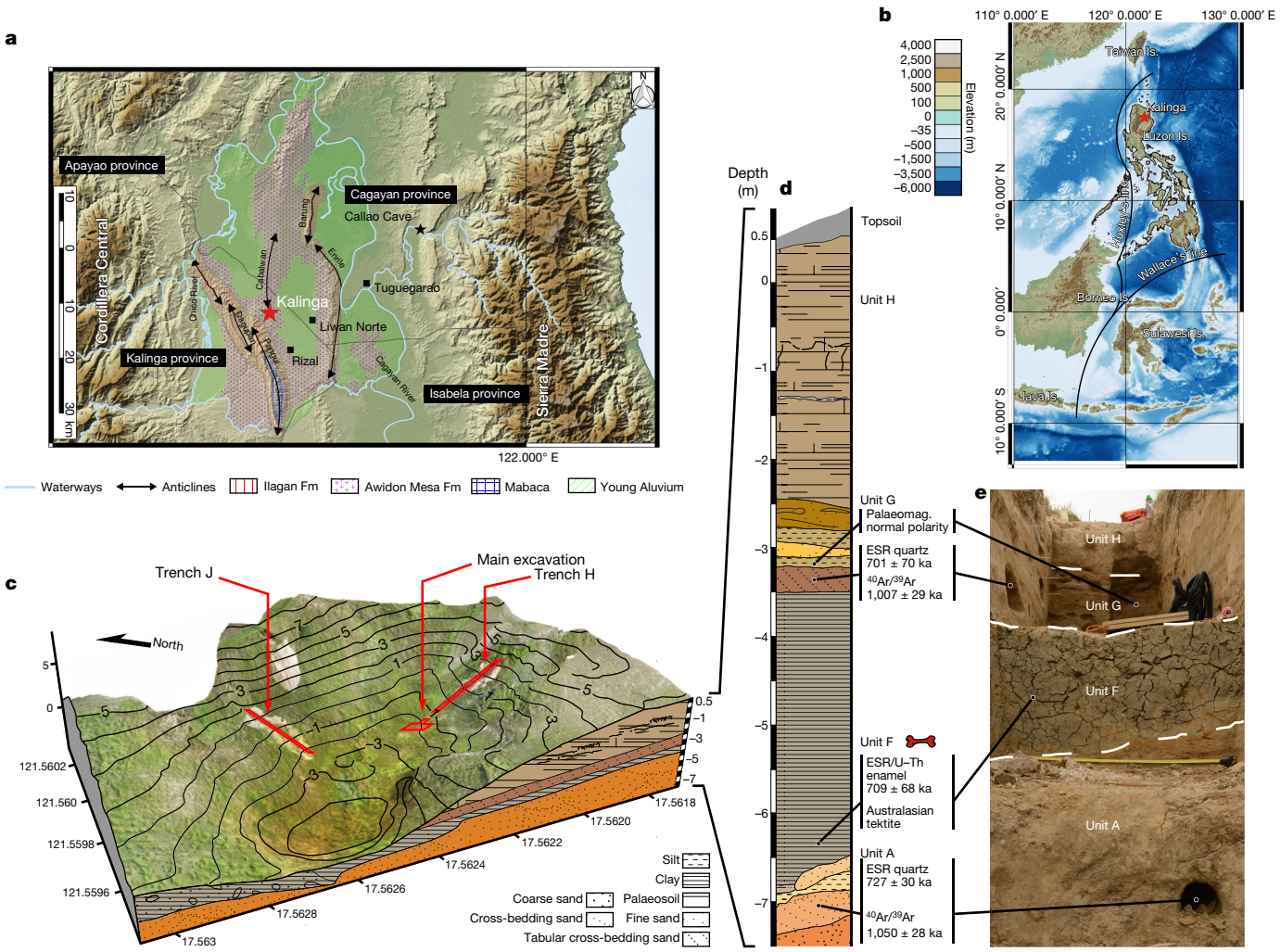


Fig. 1 | Geology and sedimentology of the Kalinga Excavation site. **a**, Digital elevation map of the Cagayan Valley surrounding the Rizal municipality (located in **b**, Northern Luzon Island, east of Huxley's and north of the Wallace Lines). The Kalinga site (red star) is located at the southern tip of the weakly folded Cabalwan Anticline. Geological units of the area bounded by the Cagayan River on the east and the Chico River on the west are after Mathisen[13]. Stratigraphically, the site layers pertain to the upper part of the Awidon Mesa Formation, a Pleistocene sequence of alluvial stream deposits intercalated with volcaniclastic and pyroclastic deposits. The depositional environment of the Awidon Mesa Formation was characterized by braided rivers on an alluvial fan system that formed in response to uplift in the Cordillera Central to the West[12,13]. **c**, Contour map of the main excavation and the adjoining trench H along the small valley where the Kalinga site is located. **d**, Detailed stratigraphy of the excavation with the absolute ages of the sedimentary units. Unit A constitutes a fining upward complex of sandy to silty cross-bedded fluvial sediments. The top of unit A is eroded and cuts down vertically over at least 2.5 m. This erosive channel is filled with unit F, a poorly sorted mudflow deposit with a maximum thickness of 3.25 m. The rhinoceros skeletal elements and most of the stone artefacts were found lying directly above the erosional contact, and were found embedded in the clay-rich mud of unit F. Unit F is conformably overlaid by a sequence of horizontally layered coarse sandy to silty layers (unit G), which is in turn conformably overlaid by a thick sequence of silty deposits overprinted by palaeosols (unit H). **e**, Southward view of trench H showing the lower and upper contacts between mudflow unit F and sandy unit G and between sandy unit A and mudflow unit F.

The Kalinga lithic assemblage is diverse in its techniques, technology and final products, and appears similar to the chert industry described at the Arubo 1 site[14] (see Supplementary Information). Also recovered from the unit F excavation area was a 600-g pebble among hundreds of pebbles that were all lighter than 200 g, and which we interpret as a possible manuport.

Among the more than 400 bones recovered from unit F, the most striking remains were of a disarticulated, approximately 75% complete skeleton of a single *R. philippinensis* individual (Fig. 3 and Extended Data Fig. 2). The bones were found lying on top of the erosional surface down-cutting unit A, and were embedded in the basal clay-rich sediments of unit F along the deepest part of the paleo-channel bed (Extended Data Fig. 3). Although none of the rhinoceros bones were found articulated, the recovered skeletal elements occur within a 3 × 2-m^2 area, suggesting that disarticulation occurred sub-aerially and that transport prior or during deposition of mudflow unit F was minimal.

Thirteen of the excavated rhinoceros bones, all of which in life had a thin cover of soft tissue (that is, the ribs and metacarpals)[15,16], display cut marks (Fig. 3 and Extended Data Fig. 2). Both rhinoceros humeri have similar percussion marks on the anterior surface for the right humerus and on the posterior surface for the left humerus, and both were presumably made with the intention to smash the bones and gain access to the marrow[17]. This percussion action resulted in the breakage of the left humerus into five pieces, which is the only bone found fragmented; however, the fragments were still clustered together within a small 1-m^2 area of the excavation. On the right humerus, however, percussion did not result in the fragmentation of the bone (Fig. 3).

To constrain the age of the bone bed and the stone artefacts it contained, we applied three different dating methods to various materials (Fig. 1). Single crystal ^{40}Ar/^{39}Ar dating was applied to plagioclase

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

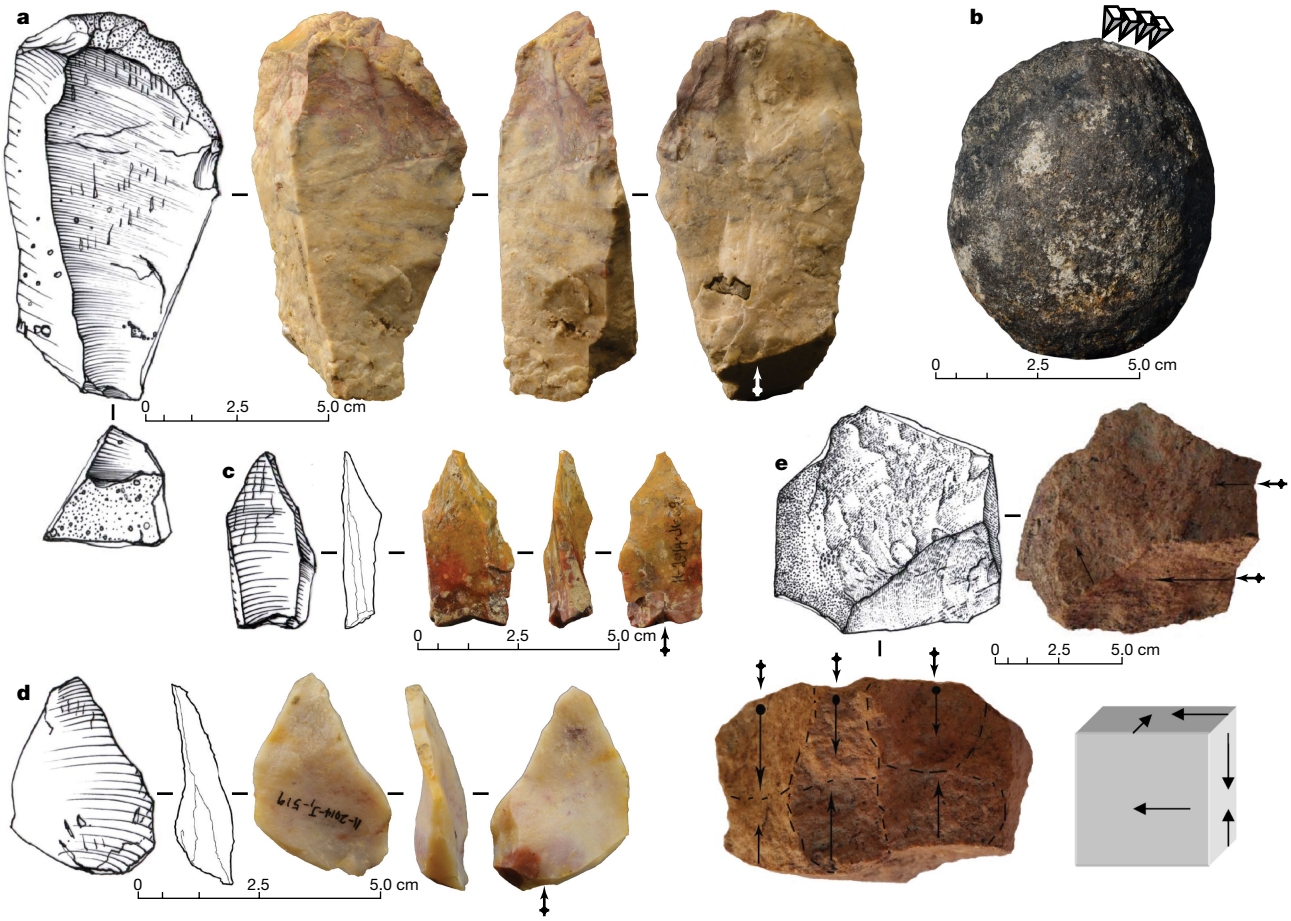


Fig. 2 | Lithic artefacts from Kalinga. a, Cortical flake on chert (II-2014-J1-362; length (L) = 100 mm, breadth (B) = 55 mm, thickness (T) = 33 mm). **b**, Possible hammerstone on dacite (II-2014-J1-371), although its highly eroded aspect precludes any definitive conclusion. Arrows indicate crushed areas interpreted as the result of precussions. **c**, Siret kombewa flake on jasper (II-2014-J1-391; L = 40 mm, B = 18 mm, T = 8 mm) that has a longitudinal and oblique fracture on the inferior two-thirds of the left side resulting from a knapping accident while flaking. **d**, Double-backed flake on flint (II-2014-J1-519). **e**, Core on quartz (II-2014-J1-396), with clear marks of knapping on an anvil, and its diachritic diagram. Arrows indicate the percusion axes.

crystals from the sandy units directly below and above the archaeological unit F and yielded two statistically undistinguishable weighted mean dates of 1,050 ± 28 ka and 1,007 ± 29 ka, respectively (1σ confidence interval; Supplementary Information and Extended Data Fig. 5). These $^{40}Ar/^{39}Ar$ dates yielded an age for the formation of the volcanic plagioclase crystals. Quartz grains from the same two sandy units were also dated using the electron spin resonance (ESR) method[18], and resulted in a maximum depositional date of 727 ± 30 ka for unit A, and a minimum depositional date of 701 ± 70 ka for unit G (1σ confidence interval; see Supplementary Information).

To directly constrain the age of the rhinoceros skeleton and the cut marks, we applied ESR/uranium-series dating to the enamel of the rhinoceros's right maxillary third premolar from the unit F bone bed. The tooth yielded an age of 709 ± 68 thousand years (1σ confidence interval), which is in agreement with the ESR results on the quartz (Fig. 1, Extended Data Fig. 6, Extended Data Table 1, Supplementary Information and Supplementary Table 1). In addition, a palaeomagnetic sample was taken from a laminated silty lens in the lower part of unit G and was found to have a normal magnetic polarity (see Supplementary Information: and Extended Data Fig. 7). The presence in unit F of a reworked Australasian tektite (see Supplementary Information and Extended Data Fig. 8) that had formed during a major meteoritic impact just before the onset of the Brunhes Normal polarity epoch at 781 ka[19,20], also provides further support for these closely grouped dating results. These results further suggest that the volcanic plagioclase crystals from unit G on which the $^{40}Ar/^{39}Ar$ date was obtained were reworked from older volcanoclastic deposits, and therefore provide a maximum age for the sequence (see Supplementary Information). Taken together with the ESR dating results, it follows that the rhinoceros skeleton was buried by a mudflow at least 631 ka.

Our excavations at Kalinga and the numeric dating results clearly provide securely dated evidence for human colonization of the Philippines by the early Middle Pleistocene epoch, and long before the appearance of modern humans in both the local context and wider Island South East Asia region[21]. Although the identity of these archaic toolmakers remains unknown, it is likely that they dispersed over at least one sea barrier to reach Luzon Island[22]. The most likely points of origin are Borneo through Palawan to the west, or China through Taiwan to the north, this latter island was connected to mainland Asia during periods with low sea levels[23]. The Middle Pleistocene fauna from the Awidon Mesa Formation contains a wider range of vertebrates than the Pleistocene faunas from two islands to the south of the Philippines that have both yielded evidence of the occupation by premodern humans, Sulawesi[9] and Flores[26] (Extended Data Fig. 9). Overseas dispersal throughout Wallacea of land mammals, including hominins, could have been primarily, although not exclusively, in a north to south direction, following the major surface current flow patterns.

Beyond the chronological gap that is yet to be filled, a question clearly linked to our discovery is the origin of the Callao Cave hominin that has been dated to 66.7 ± 1 ka. This diminutive Callao hominin may represent a direct descendent from a Pleistocene migration stock related to these early Kalinga toolmakers—similar to what happened on Flores Island—or may be derived from a more recent migration wave of anatomically modern humans[6,21,24,25].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

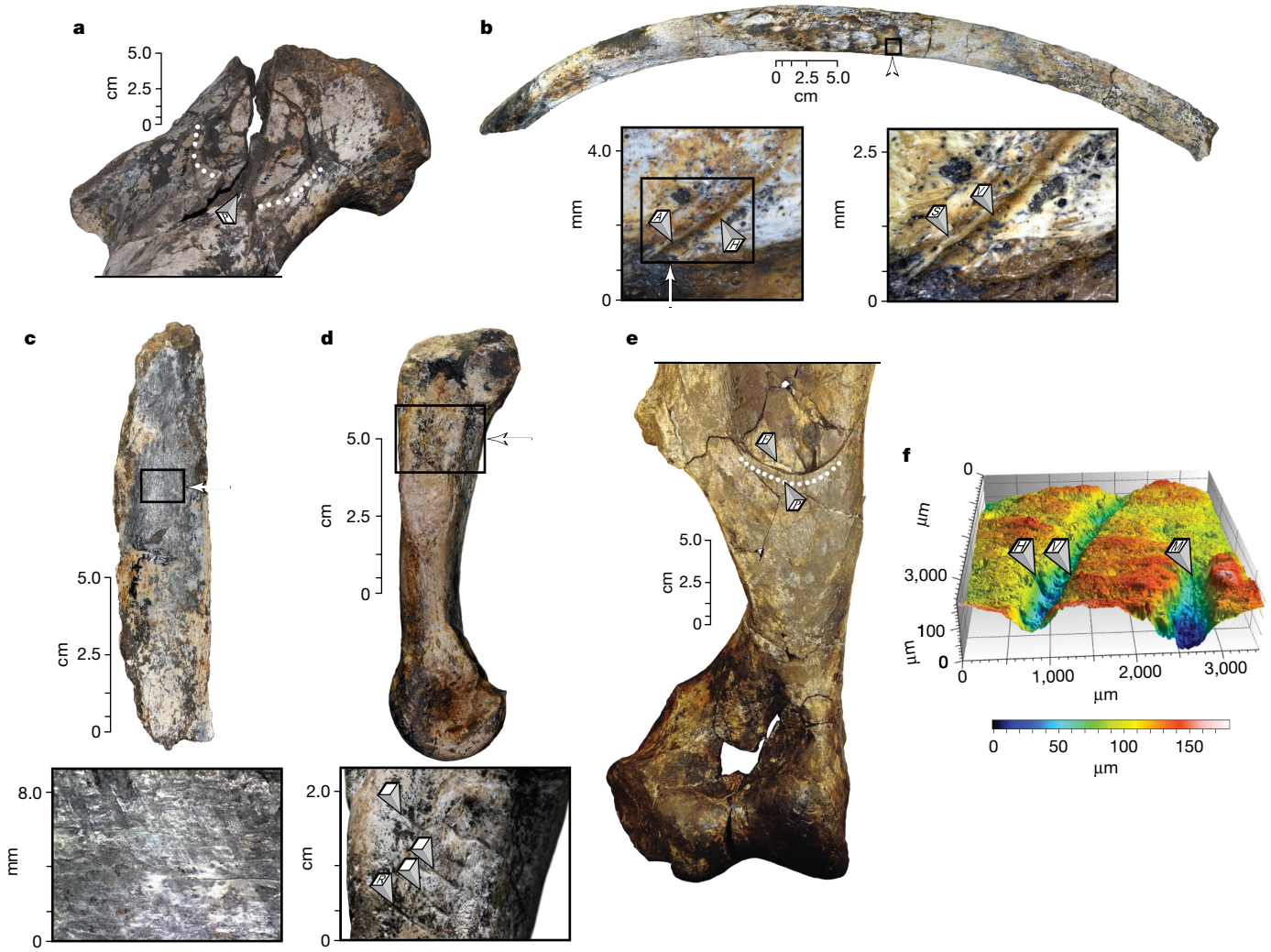


Fig. 3 | Different types of marks at the surface of the bones. a, Left humerus (II-2014-J1-368) found broken into five fragments in the excavation with an anthropogenic conchoidal percussion mark (P) on its anterior surface most likely produced to get access to the marrow. **b**, The sub-complete rib (II-2014-J1-475) has a diagnostic anthropogenic cutmark with a V-shaped cross-section (V), hertzian cones (H), asymmetrical profile (A) and shoulder effect (S) on its lateral surface resulting from defleshing. Black stains are also present inside the cutmark, which resemble the ones observed on the surface of the rib and are the result of taphonomic processes that occurred after this cutmark was made (see page 156 of Fernández-Jalvo and Andrews[17]). **c**, Rib fragment (II-2014-J1-403) with a shiny surface on its lateral face resulting from multiple multidirectional striations, which are presumably caused by trampling. **d**, Right metacarpal IV (II-2014-J1-282) with parallel and rectilinear anthropogenic cut marks (R) on its medial surface presumably generated during disarticulation. **e**, Right humerus (II-2014-J1-289) with an anthropogenic conchoidal percussion mark (P) similar in size and shape to the percussion mark on the left humerus, but located on its posterior surface and more distally and associated with a small adhered bone flake (F) (see page 298 of Fernández-Jalvo and Andrews[17]). **f**, Three-dimensional surface topography of another rib (II-2014-J1-466) showing a linear mark (on the left) with V-shaped cross-section (V) of anthropogenic origin, as well as hertzian cones (H) and a linear mark (on the right) with a base as broad as the heights of the walls of the grove, commonly attributed to trampling but also with assymetrical walls and possible microstriations in the bottom (M) of the groove, commonly attributed to anthropogenic marks.

Despite the current evidence, it still seems too farfetched to suggest that *H. erectus*, or another unknown Pleistocene ancestral candidate for the Kalinga toolmakers (for example, Denisovans[27]), were able to construct some sort of simple watercraft and deliberately cross sea barriers to reach these islands[28]. However, considering evidence of overseas dispersal during the Middle Pleistocene stage is increasing in number[29,30], such a hypothesis cannot currently be rejected.

Data availability

The datasets generated and/or analysed during the current study are available from the corresponding authors upon reasonable request.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0072-8.

Received: 10 May 2017; Accepted: 14 March 2018;
Published online 2 May 2018.

1. Van Heekeren, H. R. Early man and fossil vertebrates on the island of Celebes. *Nature* **163**, 492 (1949).
2. Maringer, J. & Verhoeven, T. Die steinartefakte aus der *Stegodon*-fossilschicht von Mengeruda auf Flores, Indonesien. *Anthropos.* **65**, 229–247 (1970).
3. von Koenigswald, G. H. R. Preliminary report on a newly-discovered Stone Age culture from northern Luzon, Philippine Islands. *Asian Perspect.* **2**, 69–70 (1958).
4. Fox, R. B. in *Paleolithic in South and East Asia* 59–85 (Mouton Publishers, The Hague, 1978).
5. Heaney, L. R. Zoogeographical evidence for Middle and Late Pleistocene land bridges to the Philippines islands. *Mod. Quat. Res. SE Asia* **9**, 127–143 (1985).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

6. Mijares, A. S. et al. New evidence for a 67,000-year-old human presence at Callao Cave, Luzon, Philippines. *J. Hum. Evol.* **59**, 123–132 (2010).
7. Ingicco, T. et al. A new species of *Celebochoerus* (Suidae, Mammalia) from the Philippines and the paleobiogeography of the genus *Celebochoerus* Hooijer, 1948. *Geobios* **49**, 285–291 (2016).
8. Brumm, A. et al. Hominins on Flores, Indonesia, by one million years ago. *Nature* **464**, 748–752 (2010).
9. van den Bergh, G. D. et al. *Homo floresiensis*-like fossils from the early Middle Pleistocene of Flores. *Nature* **534**, 245–248 (2016).
10. van den Bergh, G. D. et al. Earliest hominin occupation of Sulawesi, Indonesia. *Nature* **529**, 208–211 (2016).
11. Dizon, E. Z. & Pawlik, A. F. The lower Palaeolithic record in the Philippines. *Quat. Int.* **223–224**, 444–450 (2010).
12. Mathisen, M. & Vondra, C. The fluvial and pyroclastic deposits of the Cagayan Basin, northern Luzon, Philippines—an example of non-marine volcaniclastic sedimentation in an interarc basin. *Sedimentology* **30**, 369–392 (1983).
13. Mathisen, M. E. *Plio-Pleistocene geology of the Central Cagayan Valley, Northern Luzon, Philippines.* PhD thesis, Iowa State Univ. (1981).
14. Pawlik, A. F. The Paleolithic site of Arubo 1 in Central Luzon, Philippines. *Bull. Indo-Pacific Prehistory Assoc.* **24**, 1–10 (2004).
15. Yravedra, J. et al. Cut marks on the Middle Pleistocene elephant carcass of Áridos 2 (Madrid, Spain). *J. Archaeol. Sci.* **37**, 2469–2476 (2010).
16. Mosquera, M. et al. Barranc de la Boella (Catalonia, Spain): an Acheulean elephant butchering site from the European late Early Pleistocene. *J. Quaternary Sci.* **30**, 651–666 (2015).
17. Fernandez-Jalvo, Y. & Andrews, P. *Atlas of Taphonomic Identifications* (Springer, Dordrecht, 2016).
18. Voinchet, P. et al. ESR dating of quartz extracted from Quaternary sediments: Application to fluvial terraces system of Northern France. *Quaternaire* **15**, 135–141 (2004).
19. Lee, M.-Y., Chen, C.-H., Wei, K.-Y., Iizuka, Y. & Carey, S. First Toba supereruption revival. *Geology* **32**, 61–64 (2004).
20. Valet, J.-P. et al. Geomagnetic, cosmogenic and climatic changes across the last geomagnetic reversal from equatorial Indian Ocean sediments. *Earth Planet. Sci. Lett.* **397**, 67–79 (2014).
21. Liu, W. et al. The earliest unequivocally modern humans in southern China. *Nature* **526**, 696–699 (2015).
22. Wang, P., Li, Q. (eds). *The South China Sea: Paleoceanography and Sedimentology* Vol.13, 25–73 (Springer, Dordrecht, 2009).
23. Voris, H. K. Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *J. Biogeogr.* **27**, 1153–1167 (2000).
24. Détroit, F., Corny, J., Dizon, E. Z. & Mijares, A. S. "Small size" in the Philippine human fossil record: is it meaningful for a better understanding of the evolutionary history of the negritos? *Hum. Biol.* **85**, 45–66 (2013).
25. Sutikna, T. et al. Revised stratigraphy and chronology for *Homo floresiensis* at Liang Bua in Indonesia. *Nature* **532**, 366–369 (2016).
26. Brumm, A. et al. Age and context of the oldest known hominin fossils from Flores. *Nature* **534**, 249–253 (2016).
27. Cooper, A. & Stringer, C. B. Did the Denisovans cross Wallace's Line? *Science* **342**, 321–323 (2013).
28. Ruxton, G. D. & Wilkinson, D. M. Population trajectories for accidental versus planned colonisation of islands. *J. Hum. Evol.* **63**, 507–511 (2012).
29. Erlandson, J. M. & Fitzpatrick, S. M. Oceans, islands, and coasts: current perspectives on the role of the sea in human prehistory. *J. Island Coast. Archaeol.* **1**, 5–32 (2010).
30. Dennell, R. W., Louys, J., O'Regan, H. J. & Wilkinson, D. M. The origins and persistence of *Homo floresiensis* on Flores: biogeographical and ecological perspectives. *Quat. Sci. Rev.* **96**, 98–107 (2014).

Acknowledgements J. Barns and A. Labrador provided support for this research as well as G. Concepcion and M. dela Cruz Jr. The Kalinga excavation project was funded by the French Department for Foreign Affairs (Project MARCHE, to T.I.), The National Museum of The Philippines (to C.J.-o. and M.C.R.), The University of the Philippines Diliman Research Grant of the Office of Vice President for Academic Affairs (to T.I.) and the European Social Fund (Project ISOLARIO, NSRF Thalis-UOA, to G.L.). T.I. also received funding from a National Geographic Global Exploration grant (HJ-035R-17), from the French Centre National de la Recherche Scientifique (GDRi PalBiodivASE with Valéry Zeitoun), from Sorbonne Universités (Project MH@SU TAPHO), from the Société des Amis du Musée de l'Homme and from the LabEx BCDiv. G.D.v.d.B. received funding from an Australian Research Council (ARC) Future fellowship (FT100100384). J.d.V. received funding from the Quaternary and Prehistory Erasmus Mundus Program. Additional funding was provided by the Embassy of France to the Philippines and by the Rizal Municipality. M.G.C. received funding from CERCA Programme/Generalitat de Catalunya. We thank S. Hayes for her feedback on the manuscript; M.-M. Blanc-Valleron for providing access to the X-ray diffractometer; A. Ledoze, S. Puaud, V. Scao, S. Baillon, H. Guillou, J. Marteau, M. Bigerelle, R. Deltombe, D. Borshneck and F. Demory for their valuable help in the laboratory analyses.

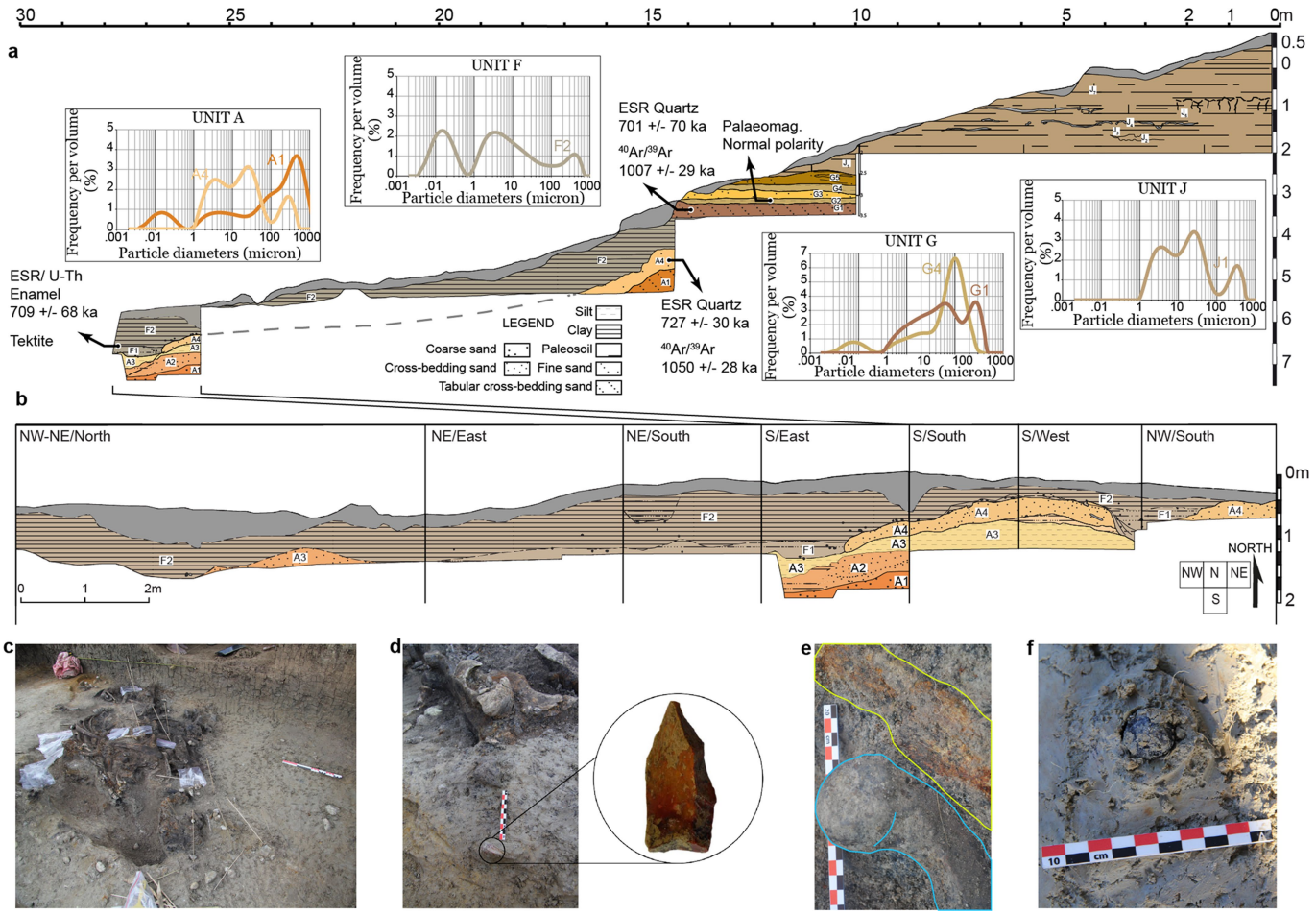
Reviewer information *Nature* thanks M. Duval, J. Kappelman, R. Potts and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.I., J.d.V. and A.B. conceived the study with C.J-o. and G.D.v.d.B. in collaboration with N.A., G.L., P.C.H.A. and M.L. The site stratigraphy was recorded and analysed by G.D.v.d.B. Samples for ESR/U-series dating were analysed by J.-J.B., Q.S. and C.F. Samples for ESR dating on quartz were analysed by P.V. Samples for $^{40}Ar/^{39}Ar$ dating were analysed by S.N. and A.P. D.Y. analysed samples for palaeomagnetism. P.R. analysed the tektite. The lithic material was studied by H.F. and M.G.C. Palaeobotanical remains were analysed by A.-M.S. and C.K. The faunal and taphonomical analysis was conducted by T.I., K.M., M.C.R. and N.A. T.I. and G.v.d.B. wrote the manuscript.

Competing interests The authors declare no competing interests.

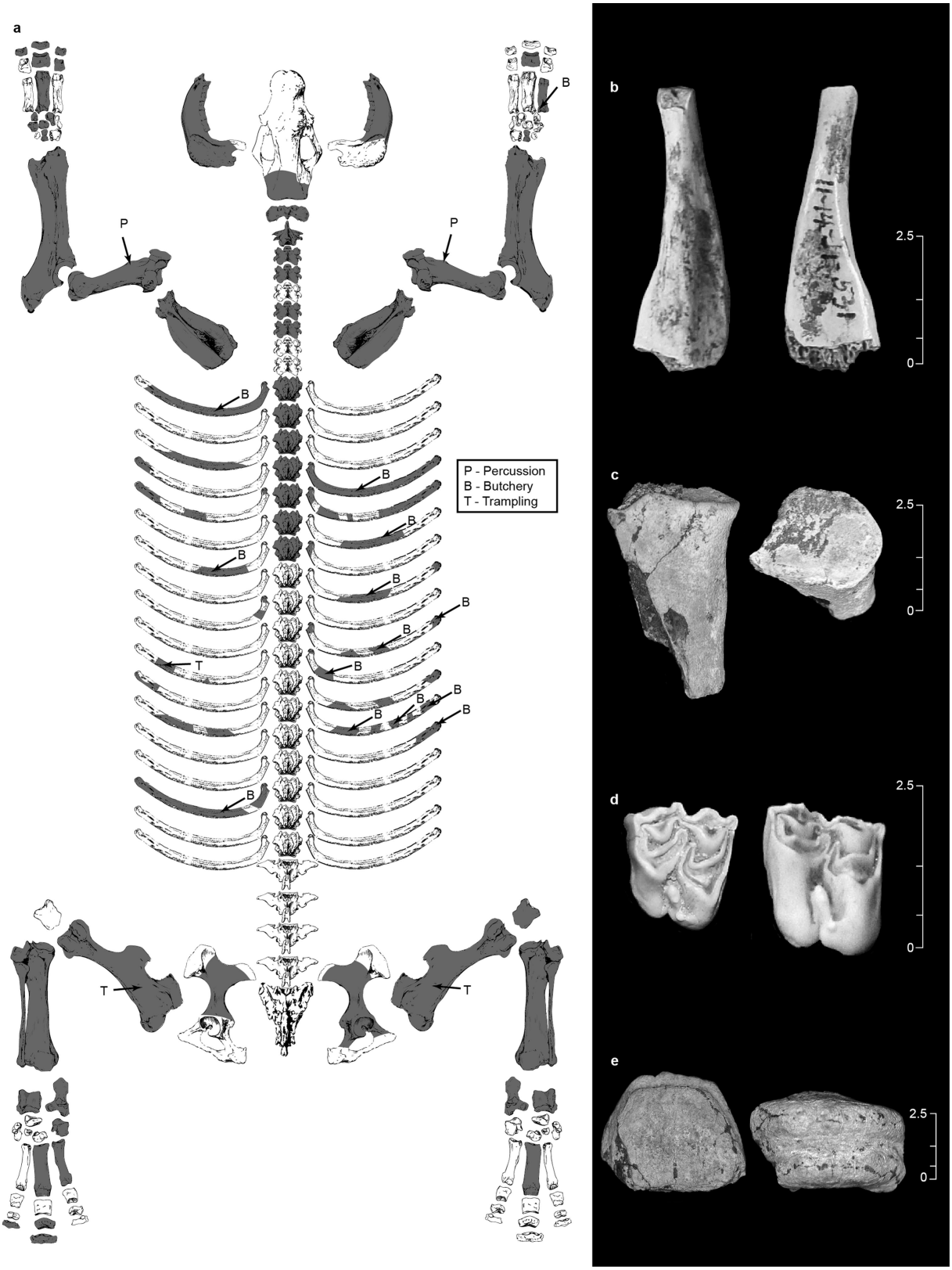
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0072-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0072-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.I. or M.C.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



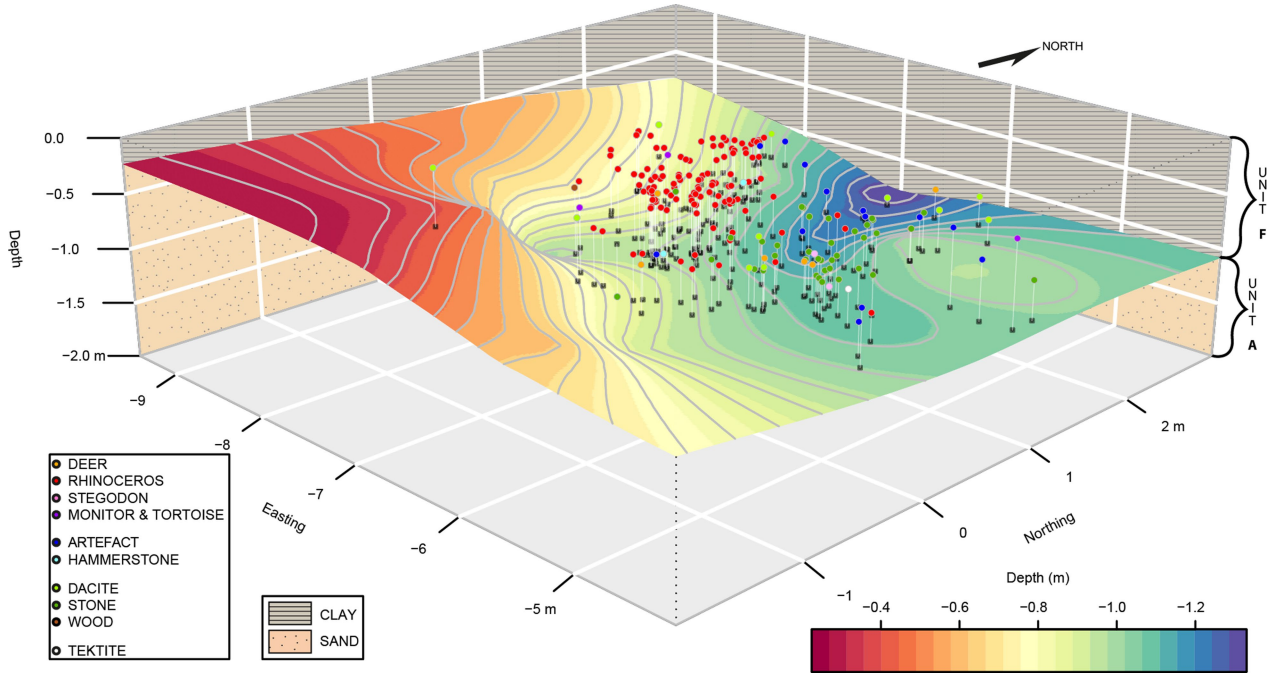
Extended Data Figure 1 | Geology and sedimentology of the Kalinga Excavation site. a, Detailed stratigraphic drawing of trench H also showing the east wall of the S quadrant of the main excavation. The sedimentary patterns are the same as in Fig. 1. Representative logarithmic grain-size diagrams are shown for samples from each sedimentary unit. **b**, Detailed stratigraphic drawing of the main excavation walls. **c**, Overview towards the northwest of the quadrants N and NW of the main excavation in 2015. The concentration of faunal remains and stone tools lying at the base of unit F is exposed, just above the eastward sloping erosional contact with unit A. **d**, Detailed view of quadrant NW showing the position of a flake lying next to the rhinoceros left femur. **e**, Detail of quadrant N showing the piece of waterlogged wood fragment (yellow outline) recovered near a rhinoceros rib extremity (blue outline). **f**, Detail of quadrant NE showing the tektite recovered in unit F along with the faunal and lithic remains.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



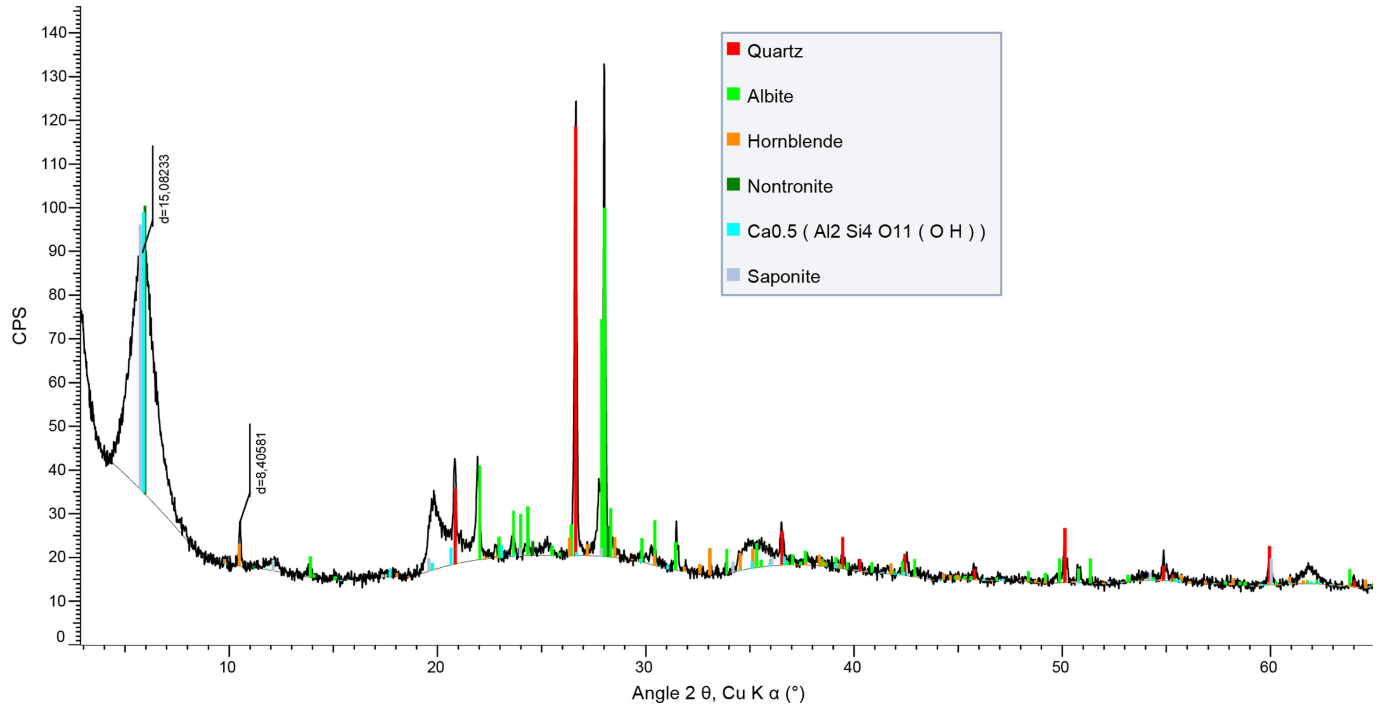
Extended Data Figure 2 | Faunal remains from Kalinga archaeological unit F. a, Drawing showing the preservation of the rhinoceros and position of the taphonomical marks. A total of 97 fragments of ribs of all sizes have been recovered and not all of them could be clearly positioned on the skeleton. We estimate that about 75% of the skeleton has been recovered. **b**, Fibula of *Varanus salvator*. **c**, Radius of a Cervidae. **d**, Molar of *Cervus* cf. *mariannus*. **e**, Molar fragment of *Stegodon* cf. *luzonensis*.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



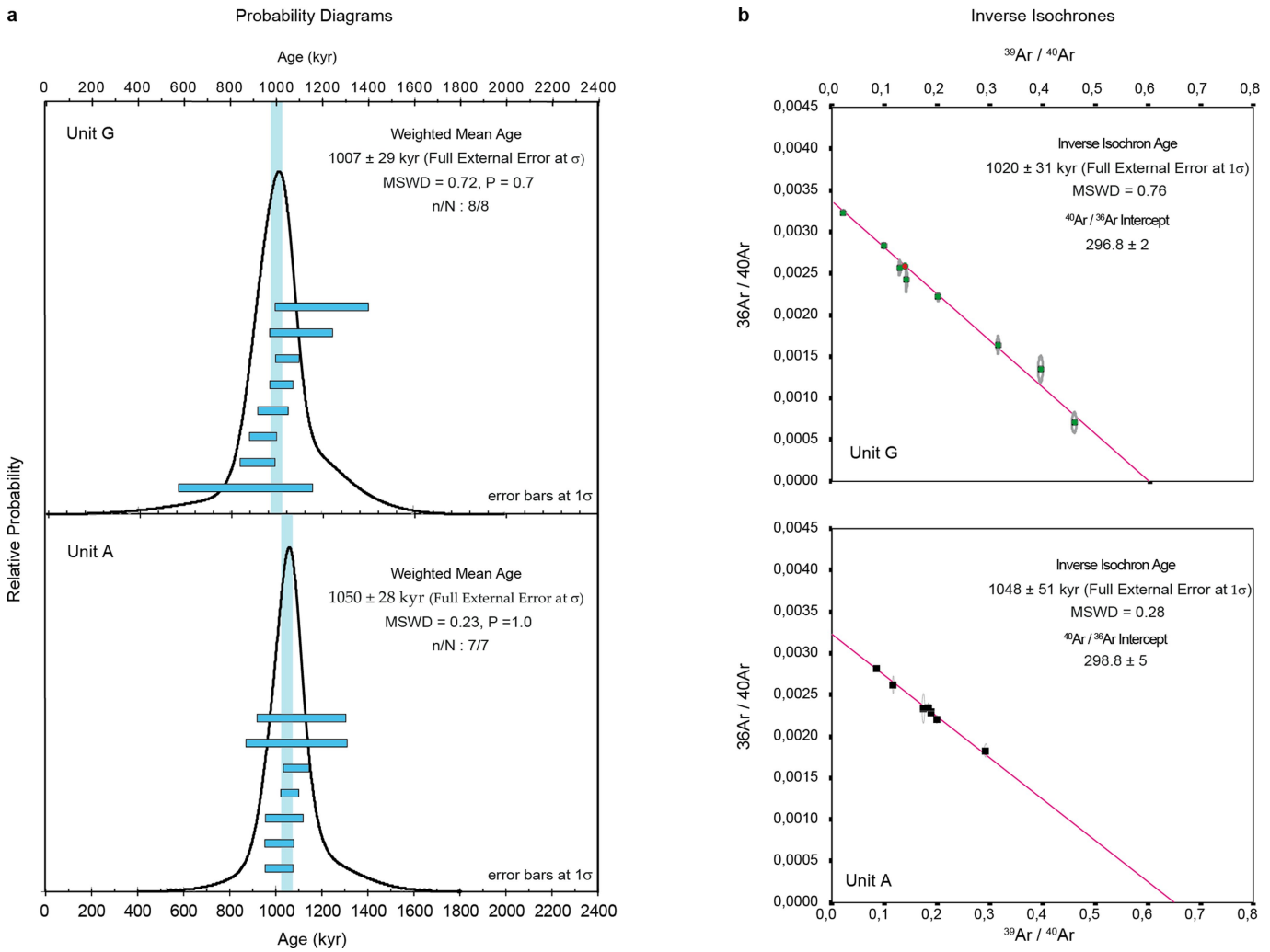
Extended Data Figure 3 | Digital elevation model of the main Kalinga excavation. The model shows the contact surface topography between unit A and unit F and the vertical projections (black squares) of the archaeological materials (coloured points) on this surface. The model was produced by interpolation through a kriging method from 37 three-dimensional coordinates recorded in the main excavation with a total station on the erosional surface that cuts down into unit A. This surface of contact corresponds to an erosional channel cutting down into the sandy unit A. All the material has been recovered lying across the base of the clay-rich unit F along this channel, between 0.7 m and 1.3 m deep.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



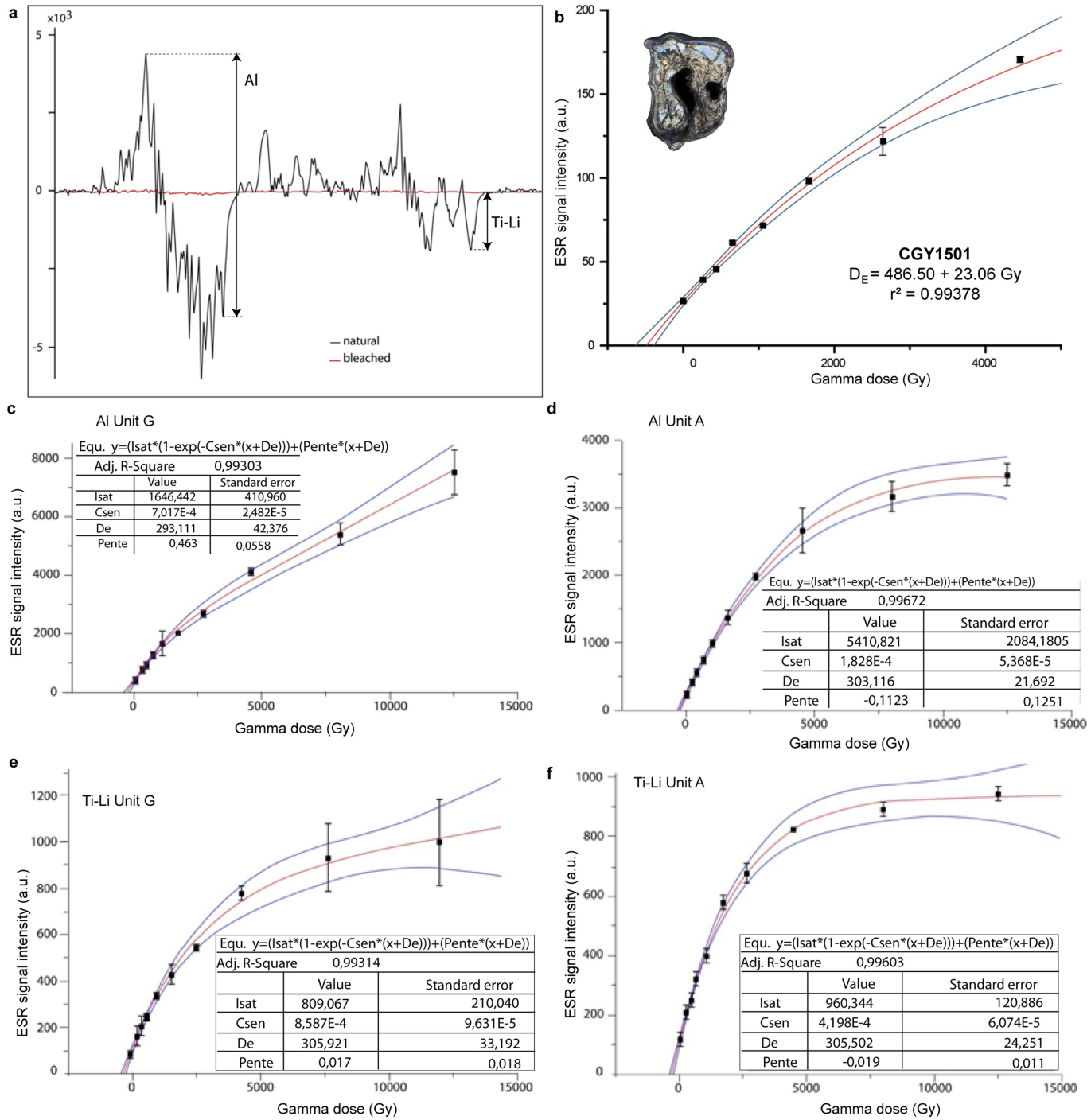
Extended Data Figure 4 | X-ray diffraction pattern of powdered unit F clays. Quartz corresponds to the bipyramidal quartz crystals, some of which are visible to the naked eye. Bipyramidal quartz, albite, hornblende, nontronite and saponite all have a volcanic origin. Albite is a plagioclase feldspar frequent in pegmatites. Hornblende is a common silicate mineral in igneous and metamorphic rocks. Nontronite is an iron-rich smectite type of clay, which can be produced by hydrothermal alteration. Similarly, the smectite mineral saponite results from alteration of volcanic glass. The mineral composition of unit F supports the interpretation as a mudflow set in a volcanic environment. CPS, counts per second; Cu K-α corresponds to the wavelength.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Figure 5 | $^{40}Ar/^{39}Ar$ fusion ages of single potassium plagioclase crystals for samples from unit A and unit G. a, **b**, Probability diagrams (**a**) are correlated to the related inverse isochrones (**b**). Individual ages in **a** are $\pm 1\sigma$.

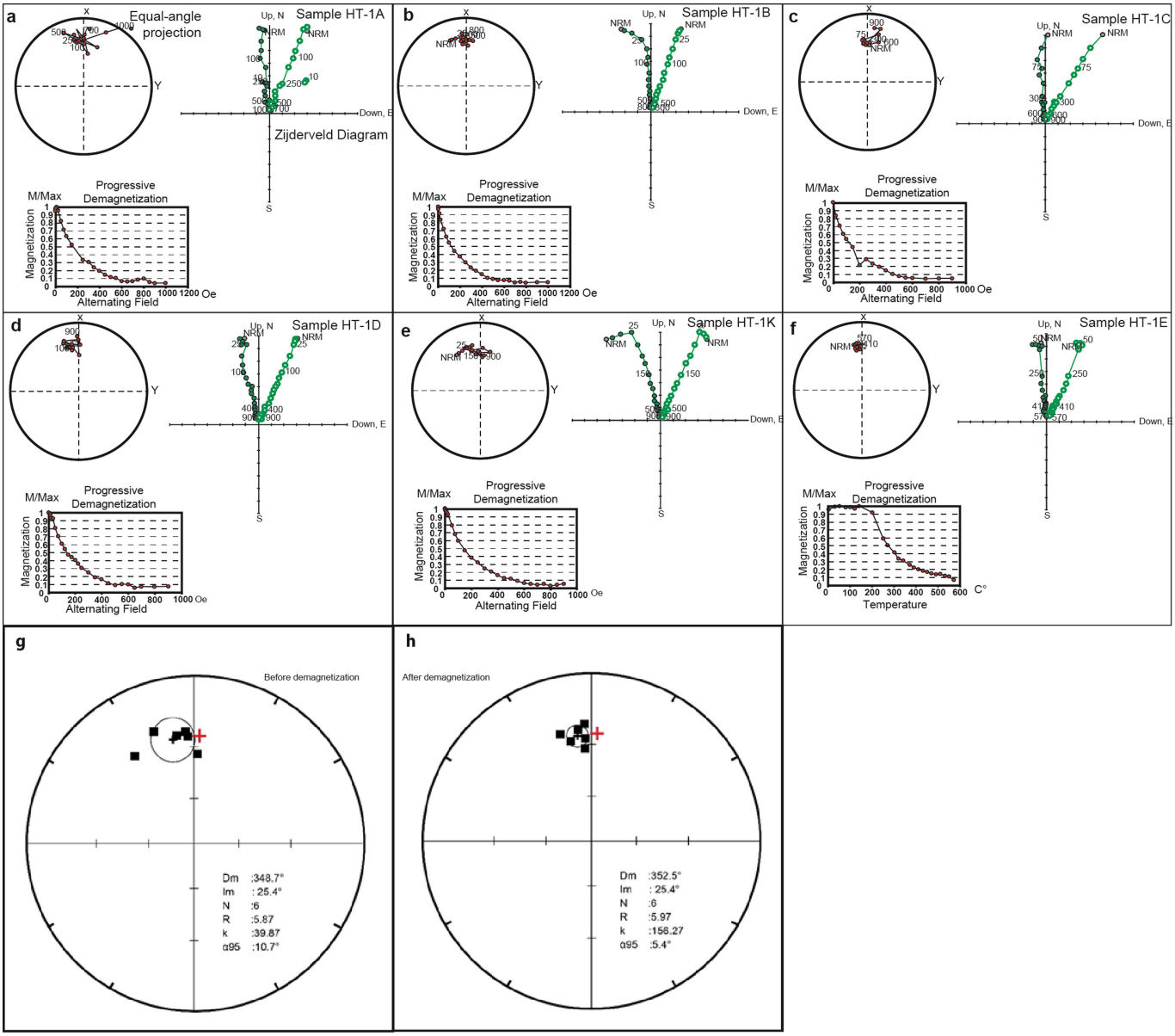
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Figure 6 | Measurements of dose rates and calculation of equivalent dose to compute the ESR ages from quartz crystals and tooth enamel. a, Al and Ti centre ESR spectra of natural and bleached aliquots for unit A quartz showing that the Al signal is not completely reset, although it is not measurable because it is extremely weak and concealed by noise. **b**, ESR dose–response curve obtained for the rhinoceros tooth (archaeological number: II-2014-J1-095; sample code: CGY1501). The equivalent dose (D_e) was extrapolated using a single saturating exponential function (Origin Microcal software) following the recommendations of Duval and Grün[31] ($D_e < 500$ Gy, therefore $D_{max} < 5{,}000$ Gy for this sample). **c**–**f**, ESR dose–response curves, for the aluminium (Al) and titanium–lithium (Ti-Li) centres of layers G1 and A4. **b**–**f**, The vertical bars indicate the standard deviation around the mean for each measurement. The red curve is the dose–response curve. The blue curves indicate the 95% confidence interval of the dose–response curve.

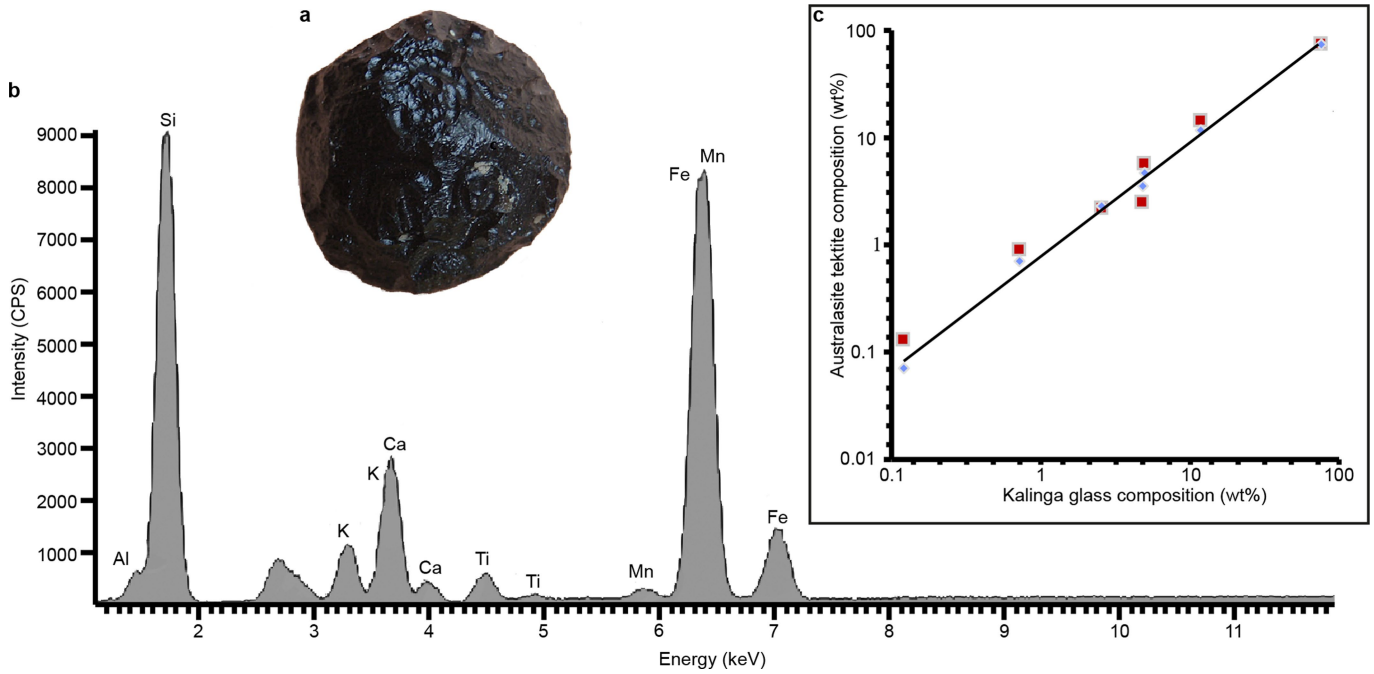
31. Duval, M. & Grün, R. Are published ESR dose assessments on fossil tooth enamel reliable? *Quat. Geochronol.* **31**, 19–27 (2016).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Figure 7 | Progressive demagnetization curves, equal area projection stereoplots and Zijderveld diagrams for the six analysed specimens from Layer G_2. a–f, Each panel shows the progressive demagnetization curves (bottom left), equal area projection stereoplots (top left) and Zijderveld diagrams (right). Open circles in the Zijderveld diagrams represent the inclination whereas closed circles represent declination. Open and closed circles in the equal area projection (stereoplot) represent the upper and lower hemisphere, respectively. **a–e**, Specimens were treated by alternating field demagnetization. **f**, Specimen HT-1E was treated by thermal demagnetization. **g**, **h**, Equal area projections of the mean chemical remanent magnetization (ChRM) directions of all analysed specimens before (**g**) and after (**h**) demagnetization. Solid squares represent the upper hemisphere. The black cross indicates the mean ChRM direction of the six specimens combined, surrounded by the α95 circle. The red cross represents the present-day magnetic direction.

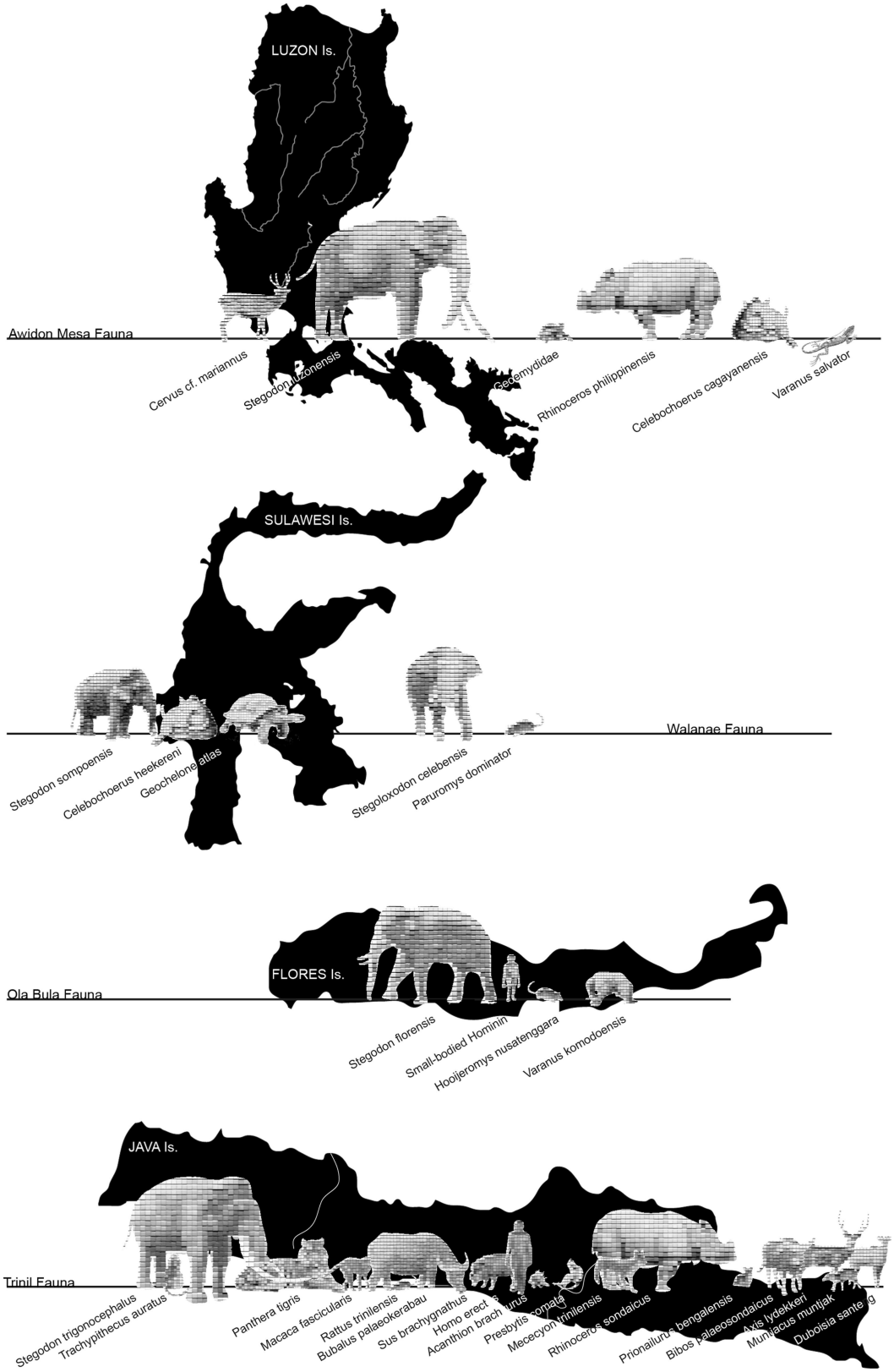
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Figure 8 | Analysis of the Kalinga tektite recovered from the archaeological layer unit F. a, Picture of the tektite. **b**, Micro-X-ray fluorescence (μXRF) spectra (un-indexed peaks correspond to the Rh source) showing its composition. **c**, Comparison of the Kalinga tektite glass composition through μXRF to an australasite tektite from China measured in the same conditions (red squares) and an average australasite composition (blue diamonds)[32].

32. Koeberl, C. The geochemistry of tektites: an overview. *Tectonophysics* **171**, 405–422 (1990).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Figure 9 | Fauna diversity of the Kalinga site compared to contemporaneous faunas from other islands in the region. Contemporaneous to those faunas are 'classic' *H. erectus* faunas on Java Island[33], the Mata Menge Fauna of Flores Island, which includes the putative ancestor of *H. floresiensis*[34], and the Walanae Fauna on the southwestern branch of Sulawesi[35,36].

33. De Vos, J., Sartono, S., Hardja-Sasmita, S. & Sondar, P.-Y. The fauna from Trinil, type locality of *Homo erectus* a reinterpretation. *Geologie Mijnbouw* **61**, 207–211 (1982).
34. van den Bergh, G. D. et al. *Homo floresiensis*-like fossils from the early Middle Pleistocene of Flores. *Nature* **534**, 245–248 (2016).
35. van den Bergh, G. D., de Vos, J. & Sondaar, P. Y. The Late Quaternary palaeogeography of mammal evolution in the Indonesian Archipelago. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **171**, 385–408 (2001).
36. Downing, K. F., Musser, G. G. & Park, L. E. in *Advances in Vertebrate Paleontology and Geochronology* Vol. 14 (eds Tomida, Y. et al.) 105–121 (National Science Museum Monographs, Tokyo, 1998).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | ESR/U-series results for the rhinoceros tooth

a

Sample	Tissue	U (ppm)	$^{234}U/^{238}U$	$^{230}Th/^{232}Th$	$^{230}Th/^{234}U$	$^{2220}Rn/^{230}Th$	Apparent U-series ages[1] (ka±2σ)
IPH-2015-09 CGY1501 (R-2014-J1-095)	dentine	52.3336 ± 0.1692	1.3701 ± 0.0062	141.40	0.8950 ± 0.0051	1.000	**220.70** ± 3.28
	enamel	0.2356 ± 0.0001	1.2564 ± 0.0009	48.96	0.8147 ± 0.0014	0.976	**164.86** ± 0.96

b

Sample	Tissue	Palaeodose D_E (Gy)	U uptake parameter p	$D_{a\,\alpha}$ Internal [1] (μGy/a)	$D_{a\,\beta}$ [2] dent+sed (μGy/a)	$D_{a\,\beta1}$ dentine (μGy/a)	$D_{a\,\beta2}$ sediment (μGy/a)	D_{a1} γ (μGy/a)	D_{a2} cosmic (μGy/a)	D_a total (μGy/a)	ESR/U-series ages US [4] (ka±1σ)
IPH-2015-09 CGY1501 (R-2014-J1-095)	dentine	486.50 ± 23.06	-0.1636 ± 0.0447	47 ± 8	258 ± 35	208 ± 35	50 ± 4	242 ± 12	145 ± 10	692 ± 38	**709 ± 58**
	enamel		0.3834 ± 0.0765								

a, U-series data for the tooth CGY1501. **b**, ESR/U-series data and age. Eventual radium and radon losses from the dental tissues were estimated from cross-checked γ and α data[37]. Dose conversion factors were used according to previously published data[38]. Water contents of 0 ± 0%, 7 ± 5% and 10 ± 5% were used for enamel (fixed), dentine (fixed) and sediments (difference in mass between the natural sample and the same sample dried for a week in an oven at 50 °C), respectively.

[1]A *k*-value of 0.13 ± 0.02 was used following a previously published study[38].

[2]The enamel thickness removed during the preparation process was taken into account in the β contribution calculation[39]. The following values were used for the age calculation: 1,505 ± 376 μm, 351 ± 88 μm and 45 ± 12 μm for the initial thickness, thickness after preparation on the dentine side and on the sediment side, respectively.

[3]The cosmic dose was estimated based on the depth using Prescott and Hutton's formulae[40]. Because we have at present no means to know precisely when the erosion took place and since when the archaeological material became buried under less than 7 m of sediment, a depth of 2.75 m was used for the cosmic dose rate estimation as an intermediate value between the 7 m of sediments that once covered the archaeological layer and the present 70 cm to 1.20 m depth at which the archaeological material was recovered, and as an average between the once full thickness of the archaeological layer and the present thickness from which the archaeological material was recovered. A cosmic dose estimated from a depth of 7 m would result in a 10% older age for unit F and a cosmic dose estimated from a depth of 1 m would result in a 7% younger age for unit F.

[4]Uncertainties on the ESR/U-series ages were calculated using Monte Carlo approach[41].

37. Bahain, J.-J., Yokoyama, Y., Falguères, C. & Sarcia, M. N. ESR dating of tooth enamel: a comparison with K–Ar dating. *Quat. Sci. Rev.* **11**, 245–250 (1992).
38. Grün, R. & Katzenberger-Apel, O. An alpha irradiator for ESR dating. *Anc. TL* **12**, 35–38 (1994).
39. Brennan, B. J., Rink, W. J., McGuirl, E. L., Schwarcz, H. P. & Prestwich, W. V. Beta doses in tooth enamel by "one-group" theory and the ROSY ESR dating software. *Radiat. Meas.* **27**, 307–314 (1997).
40. Prescott, J. R. & Hutton, J. T. Cosmic ray contributions to dose rates for luminescence and ESR dating: large depths and long-term time. *Radiat. Meas.* **23**, 497–500 (1994).
41. Shao, Q., Bahain, J.-J., Dolo, J.-M., Falguères, C. Monte Carlo approach to calculate US-ESR ages and their uncertainties. *Quat. Geochronol.* **22**, 99–106 (2014).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0074-6

Lethal trap created by adaptive evolutionary response to an exotic resource

Michael C. Singer[1,2]* & Camille Parmesan[1,2,3]*

Global transport of organisms by humans provides novel resources to wild species, which often respond maladaptively. Native herbivorous insects have been killed feeding on toxic exotic plants, which acted as 'ecological traps'[1–4]. We document a novel 'eco-evolutionary trap' stemming from the opposite effect; that is, high fitness on an exotic resource despite lack of adaptation to it. *Plantago lanceolata* was introduced to western North America by cattle-ranching. Feeding on this exotic plant released a large, isolated population of the native butterfly *Euphydryas editha* from a longstanding trade-off between maternal fecundity and offspring mortality. Because of this release—and despite a reduced insect developmental rate when feeding on this exotic—*Plantago* immediately supported higher larval survival than did the insects' traditional host, *Collinsia parviflora*[5]. Previous work from the 1980s documented an evolving preference for *Plantago* by ovipositing adults[6]. We predicted that if this trend continued the insects could endanger themselves, because the availability of *Plantago* to butterflies is controlled by humans, who change land management practices faster than butterflies evolve[6]. Here we report the fulfilment of this prediction. The butterflies abandoned *Collinsia* and evolved total dependence on *Plantago*. The trap was set. In 2005, humans withdrew their cattle, springing the trap. Grasses grew around the *Plantago*, cooling the thermophilic insects, which then went extinct. This local extinction could have been prevented if the population had retained partial use of *Collinsia*, which occupied drier microhabitats unaffected by cattle removal. The flush of grasses abated quickly, rendering the meadow once again suitable for *Euphydryas* feeding on either host, but no butterflies were observed from 2008 to 2012. In 2013–2014, the site was naturally recolonized by *Euphydryas* feeding exclusively on *Collinsia*, returning the system to its starting point and setting the stage for a repeat of the anthropogenic evolutionary cycle.

In his 1999 address to the Ecological Society of America, Gary Polis described *Homo sapiens* as a 'ubiquitous keystone pest', and so we have proven to be[7]. Although adaptation to human activities enables some wild species to coexist with us[8], many fail to adapt to human land management and suffer in consequence. An example of this is the setting by humans of ecological traps[9,10], which are defined as follows: "in an environment altered suddenly by human activities, an organism makes maladaptive habitat choices based on formerly reliable environmental cues, despite availability of higher quality habitat"[10].

Most ecological traps result from preference by wild organisms for novel resources that are unsuitable or toxic. Australian monitor lizards suffered population crashes after feeding on toxic exotic cane toads[11]. Bees have maladaptively preferred crops grown from seed treated with neonicotinoid insecticides[12]. Several examples have involved insect herbivores feeding on exotic host plants that reduced insect fitness[1–4] but fell short of causing population extinctions, because traditional hosts were still used alongside the exotics.

In the ecological trap scenario, the novel resource is accepted as food but is initially detrimental[10], with the expectation that evolution should lead either to behavioural avoidance of the resource or physiological ability to use it[3]. Here we describe a different type of trap formed by a novel resource that immediately supported such high fitness that a butterfly population evolved complete dependence on it, causing local extinction when humans withdrew our apparent gift. This paradoxical phenomenon is previously undescribed and does not fit published definitions of either ecological trap or evolutionary trap[1,10]. We here describe it as an eco-evolutionary trap.

Our study insects—the thermophilic[13] Edith's checkerspot butterflies, *E. editha*—formed a sedentary, isolated population around the margins of a spring-fed meadow (Schneider's Meadow; hereafter 'Schneider') at 1,800 m elevation in Carson City, Nevada. This species has one generation per year. The novel host of this population was the exotic perennial, *P. lanceolata*, and their traditional host was the ephemeral native annual, *C. parviflora* (Extended Data Fig. 1).

The geographically closest *E. editha* populations of the same ecotype as Schneider, but where *Plantago* had not arrived, used *Collinsia* as their sole host[5]. When neonate larvae from one of these 'ancestral' populations were transplanted to Schneider, their survival on *Plantago* was identical to that of the local Schneider insects[5]. The ancestral populations were ready to use *Plantago* from the moment of its introduction. It is not surprising, then, that this exotic has been colonized twice by other North American *Euphydryas*[14,15]. Oregon *E. editha taylori* are now dependent on the exotic, although it is not clear if this is due to evolution of the butterfly, because the original host(s) have disappeared[15].

During the 1980s, survival of *E. editha* at Schneider was consistently higher on the exotic than on the traditional host (Table 1), despite larval growth being about 18% slower on *Plantago*[5]. In situations in which the principal host is ephemeral (as is *Collinsia*), female *E. editha* face a trade-off between maternal fecundity and offspring survival. Prolonging larval development can increase fecundity but the resulting delay in adult emergence augments the risk of offspring mortality from host senescence[16]. The evolutionary response to this trade-off has been to delay emergence to the point that many offspring routinely starve from phenological asynchrony with their hosts[16]. The stage is set for the time constraint to be released and fitness increased by host-switching to the longer-lived *Plantago*, despite slower larval growth on it. Indeed, the majority of larval mortality observed on *Collinsia* was from host senescence, whereas *Plantago* did not senesce during the seasons of larval activity.

No adult females sampled from ancestral populations preferred *Plantago* over *Collinsia* for oviposition, but around 20% accepted both hosts equally[5]. By contrast, by 1982 *Plantago* was already preferred for oviposition by a minority of preference-tested adults at Schneider[6]. Evolution of *Plantago* preference had begun.

Given natural selection for oviposition on *Plantago*, and given that oviposition preferences at Schneider were both heritable (estimated heritability 0.9) and correlated with offspring performance[17], we expected to see rapid evolution of preference, which did indeed occur: the proportion of insects preferring *Plantago* for oviposition increased from about 7% to around 50% by 1990[6]. This change was heritable.

[1]Biological and Marine Sciences, University of Plymouth, Plymouth, UK. [2]Theoretical and Experimental Ecology Station, CNRS, Université Paul Sabatier, Moulis, France. [3]Department of Geological Sciences, University of Texas at Austin, Austin, TX, USA. *e-mail: michael.singer@plymouth.ac.uk; camille.parmesan@plymouth.ac.uk

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Table 1 | Survival of *E. editha* placed in the field on *Collinsia* and *Plantago* at Schneider's Meadow in the 1980s

Year and reference	Life stages measured	Group survival *Collinsia*	Group survival *Plantago*	Individual survival *Collinsia*	Individual survival *Plantago*	Statistical significance (group survival)
1980 (previously unpublished)	Oviposition to second instar	33% ($n=15$)	70% ($n=20$)	Not recorded	Not recorded	$P=0.044$
1982[29]	Oviposition to second instar	17% ($n=58$)	55% ($n=86$)	4% ($n=1,810$)	24% ($n=2,764$)	$P=0.000011$
1985[5,17]	Larval survival for ten days from hatching	62% ($n=63$)	84% ($n=62$)	19.9% ($n=1,260$)	27.1% ($n=1,240$)	$P=0.0085$ (combining data from both references)
1986 (previously unpublished)	Oviposition to second instar	29% ($n=28$)	80% ($n=25$)	Not recorded	Not recorded	$P=0.00028$

Proportions of groups surviving analysed as 2×2 contingency tables by two-tailed Fisher's exact test.

Laboratory-raised, *Collinsia*-fed offspring of butterflies caught in the field in 1990 were significantly more *Plantago*-preferring than were similarly raised offspring of insects caught in the field in 1983[6].

Here we report that this bout of anthropogenic evolution continued until monophagy on *Plantago* was achieved. In 2005 and 2007, all tested females preferred to oviposit on the exotic (Fig. 1a) and in 2007 all larvae found in the field were on *Plantago* (Fig. 1b). The insects had abandoned both their traditional host (*Collinsia*) and the minor host *Penstemon rydbergii*, which had been incorporated into their diet during the host shift (Extended Data Table 1 and Extended Data Fig. 2).

In 1993, we wrote that this episode of anthropogenic evolution was "foreshadowing a new problem in conservation biology. By adapting genetically to human-induced changes, the insects risk becoming dependent on continuation of the same practices. This is a serious risk, because human cultural evolution can be even faster than the rapid genetic adaptation that the insects can evidently achieve"[6].

This prediction was fulfilled. In late 2005, following the death of Harry Schneider, the meadow was sold and cattle-grazing ceased. Grasses then grew freely, and by March 2007 96% of *Plantago* plants had become embedded in grass (Table 2 and Extended Data Figs. 3, 4a); *E. editha* larvae wandered among dense vegetation, no longer able to bask in sunlight on bare ground adjacent to their hosts.

The conspicuous communal webs spun by gregarious young larvae render *E. editha* easy to census (Extended Data Fig. 4b) and population extinction feasible to establish. Searches for larval webs, eggs and adults found no *E. editha* in 2008, 2009, 2010 or 2012 (Fig. 1c and Extended Data Table 1): the Schneider population was extinct.

By analogy with known cause–effect relationships involving other thermophilic butterflies, we attribute this extinction to the flush of lush vegetation caused by cattle removal. Lushness is associated with high rates of predation on butterfly larvae[18] and increased lushness caused by the abandonment of traditional land management in Europe has caused ground-level cooling, which has resulted in butterfly population declines and local extinctions[19–21]. The extinction of the large blue butterfly (*Phengaris arion*) in the UK has been attributed to microclimatic cooling caused by increased lushness, which itself followed a reduction of grazing by rabbits after myxomatosis[21]. The restoration of grazing and re-warming of ground-level microclimate were essential for the successful re-introduction of this butterfly[21].

Particularly in sparse vegetation, sunshine creates thermal stratification with microclimates that are hotter close to the ground[22], which speeds insect development[23]. We used previous observations of natural oviposition sites at Schneider to measure 'eggspace' temperatures. Eggspaces on exposed *Plantago* plants were augmented by 13.4 °C above ambient temperature, compared to 6.0 °C on embedded *Plantago* plants; if placed on embedded plants, eggs would be on average >7 °C cooler than if placed on exposed plants (Extended Data Table 2). In light of the high proportion of plants embedded in 2007 (Table 2), and of the known adverse effects of host embedding[15,18] and microclimatic cooling[19–21] on other butterflies, it is not surprising that this cooling was followed by extinction.

As anthropogenic nutrients were used up, the flush of grasses abated naturally in 2008; since this time, *Plantago* plants exposed to full

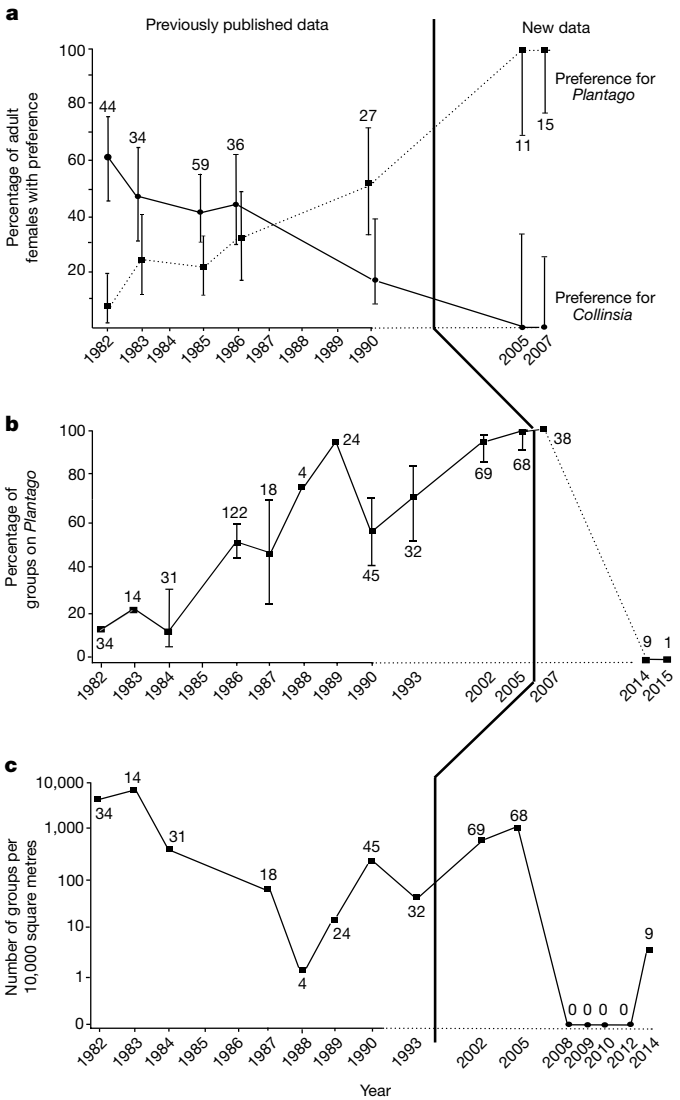


Fig. 1 | Changes in *E. editha* oviposition preference, larval diet and population density. a, Changes in proportions of adults preferring to oviposit on *Plantago* and *Collinsia*. The two graphs are not mirror images because insects without preference are not shown. Changes in strength of preference are shown in Extended Data Fig. 7. **b**, Changes in larval diet. The absence of larvae on *Collinsia* in 1988–1989 was caused by temporary spatial restriction during and after a bottleneck; in these two years, larvae were restricted to *Plantago* and *Penstemon* (Extended Data Table 1 and Extended Data Fig. 2). **c**, Changes in larval density. For all three panels numbers within graphs represent biologically independent sample sizes; individual adult insects (**a**) or larval groups (**b**, **c**). Error bars = 95% confidence intervals calculated according to Newcombe[30], with continuity correction; source data are given in Extended Data Table 3. No error bars are given for larval diet in years of inadequate sampling (1982, 1983 and 2015) or total census (1988,1989 and 2014).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Table 2 | Estimated host densities and percentages of each host embedded and exposed

Year	*Plantago* density (plants per m^2)	*Collinsia* density (plants per m^2)	Percentage of *Plantago* embedded	Percentage of *Collinsia* embedded
1984	2.3 (50 in 244 30-cm^2 quadrats)	103 (2,266 in 244 30-cm^2 quadrats)	<10%	<5%
2005	9.4 (317 in 15 2.25-m^2 quadrats)	27.5 (330 in 12 1-m^2 quadrats)	<10%	<5%
2007	0.21 (18 in 85 1-m^2 quadrats)	35.2 (528 in 15 1-m^2 quadrats)	96.4% (189/196)	10.5% (69/657)
2008	2.8 (44 in 16 1-m^2 quadrats)	26.1 (418 in 16 1-m^2 quadrats)	18.6% (43/231)	13.8% (77/559)
2009	2.5 (48 in 19 1-m^2 quadrats)	45.7 (869 in 19 1-m^2 quadrats)	11.8% (251/2,115)	11.7% (114/975)
2014	3.4 (173 in 50 1-m^2 quadrats)	52.4 (472 in 9 1-m^2 quadrats)	8.5% (28/328)	18.0% (150/833)

sunlight have again been available (Table 2 and Extended Data Fig. 5). In 2013–2014, butterflies recolonized: an exhaustive search in 2014 revealed nine *E. editha* larval webs, all of which were on the ancestral host *Collinsia* (Fig. 1b, c and Extended Data Fig. 4b). Recolonization had occurred from a population that resembled the starting condition, before the anthropogenic evolution.

The distance from Schneider's Meadow to the nearest known population of the subspecies *E. editha monoensis* (at Simee Dimeh summit) is 37.7 km. How far is this to a *Euphydryas*? Previous[24] measurements of the colonization of empty habitat patches by the bay checkerspot (*E. editha bayensis*) found that the greatest cumulative distance travelled in ten years was 4.5 km. Given the sedentary nature of the butterfly and the physical isolation of the meadow, we did not consider the possibility of rapid recolonization. Our mindset in 2014 was simply to reconfirm the extinction, and we were very surprised to find larvae. We later discovered that the 'Carter Springs' fire in September 2012 had positively affected the butterflies of Simee Dimeh by extending the size and lifespan of *Collinsia* (Extended Data Fig. 6), and causing a population boom of *E. editha* similar to a previously documented response to fire[25]. This provides a plausible source for the recolonizing *Collinsia*-feeding butterflies, despite the distance they would have had to travel.

Could the original population have survived if it had retained its traditional diet of *Collinsia* alongside *Plantago*, as it did from 1982–1987 and 1990–2002? *Collinsia* was both most abundant and most used by the insects in dry sagebrush around the meadow edge (Extended Data Figs. 1–3), where removal of cattle did not result in embedding of *Collinsia* plants, even at peak lushness in 2007 (Table 2 and Extended Data Fig. 3). If the butterflies had adopted the exotic less completely, they would probably have survived the change in land use. Conversely, if they had remained monophagous on *Collinsia* they may not have survived the bottleneck in 1988–1989[6], when—after record-breaking frost without insulating snow (−25 °C at Minden on 1 January 1988)—the population was spatially restricted to a small, sheltered, south-facing area from which *Collinsia* was coincidentally absent (Fig. 1b, c; see discussion of this event in Extended Data Fig. 2).

The evolution of *E. editha* at Schneider illustrates the process by which, over thousands of years, European grassland butterflies evolved widespread dependence on human haymaking and grazing, rendering them vulnerable to the abandonment of traditional management techniques[19–21]. However, *E. editha* as a species is not threatened by the eco-evolutionary trap that we document. The ecotypic variation and rapid evolution of this species augur well for its resilience to environmental fluctuations, whether natural or anthropogenic[22]. By contrast, substantial perturbation occurred at the subspecies level, as *E. e. monoensis* is currently restricted to four known sites: two isolated populations and two metapopulations distributed along 235 km of the eastern Sierra Nevada from McGee Creek at latitude 37° 18′ 24″ N to Schneider at latitude 39° 6′ 33″ N.

Unless the Schneider population sent out successful propagules, which we judge to be unlikely, the lineage that we observed from 1982 to 2007 is extinct. At the population level, the changes that we observed exemplify marked, oscillating anthropogenic evolution of a species not directly targeted by humans. This example of small-scale oscillating diet evolution oddly mimics the repeated recolonizations of abandoned hosts detected across millions of years by phylogenetic analysis of the butterfly family Nymphalidae, to which *E. editha* belongs[26].

Attempts at the 'evolutionary rescue' of wild species[27,28] may be compromised when anthropogenic traps such as the one documented here remain undetected. These traps may be cryptic to humans, but understanding them may become increasingly important to species conservation in the Anthropocene.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0074-6.

Received: 23 August 2017; Accepted: 24 March 2018;
Published online 9 May 2018.

1. Schlaepfer, M. A., Sherman, P. W., Blossey, B. & Rynge, M. C. Introduced species as evolutionary traps. *Ecol. Lett.* **8**, 241–246 (2005).
2. Nakajima, M., Boggs, C. L., Bailey, S., Reithel, J. & Paape, T. Fitness costs of butterfly oviposition on a lethal non-native plant in a mixed native and non-native plant community. *Oecologia* **172**, 823–832 (2013).
3. Morton, T. A. L. et al. Modeling the decline and potential recovery of a native butterfly following serial invasions by exotic species. *Biol. Invasions* **17**, 1683–1695 (2015).
4. Yoon, S. & Read, Q. Consequences of exotic host use: impacts on Lepidoptera and a test of the ecological trap hypothesis. *Oecologia* **181**, 985–996 (2016).
5. Thomas, C. D. et al. Incorporation of a European weed into the diet of a North American herbivore. *Evolution* **41**, 892–901 (1987).
6. Singer, M. C., Thomas, C. D. & Parmesan, C. Rapid human-induced evolution of insect–host associations. *Nature* **366**, 681–683 (1993).
7. Hendry, A. P., Gotanda, K. M. & Svensson, E. I. Human influences on evolution, and the ecological and societal consequences. *Phil. Trans. R. Soc. Lond. B* **372**, 20160028 (2017).
8. Thomas, C. D. *Inheritors of the Earth: How Nature is Thriving in an Age of Extinction* (Penguin/Allen Lane Books, London, 2017).
9. Robertson, B. A., Rehage, J. S. & Sih, A. Ecological novelty and the emergence of evolutionary traps. *Trends Ecol. Evol.* **28**, 552–560 (2013).
10. Schlaepfer, M. A., Runge, M. C. & Sherman, P. W. Ecological and evolutionary traps. *Trends Ecol. Evol.* **17**, 474–480 (2002).
11. Jolly, C. J., Shine, R. & Greenlees, M. J. The impacts of a toxic invasive prey species (the cane toad, *Rhinella marina*) on a vulnerable predator (the lace monitor, *Varanus varius*). *Biol. Invasions* **18**, 1499–1509 (2016).
12. Kessler, S. et al. Bees prefer foods containing neonicotinoid pesticides. *Nature* **521**, 74–76 (2015).
13. Weiss, S. B., Murphy, D. D. & White, R. R. Sun, slope, and butterflies: topographic determinants of habitat quality for *Euphydryas editha*. *Ecology* **69**, 1486–1496 (1988).
14. Brown, L. M., Breed, G. A., Severns, P. M. & Crone, E. E. Losing a battle but winning the war: moving past preference–performance to understand native herbivore–novel host plant interactions. *Oecologia* **183**, 441–453 (2017).
15. Severns, P. M. & Warren, A. D. Saving an imperiled butterfly, *Euphydryas editha taylori* (Taylor's checkerspot), by selectively conserving and eliminating exotic plants. *Anim. Conserv.* **11**, 476–483 (2008).
16. Singer, M. C. & Parmesan, C. Phenological asynchrony between herbivorous insects and their hosts: signal of climate change or pre-existing adaptive strategy? *Phil. Trans. R. Soc. Lond. B* **365**, 3161–3176 (2010).
17. Singer, M. C., Ng, D. & Thomas, C. D. Heritability of oviposition preference and its relationship to offspring performance within a single insect population. *Evolution* **42**, 977–985 (1988).
18. Wiklund, C. & Friberg, M. Enemy-free space and habitat-specific host specialization in a butterfly. *Oecologia* **157**, 287–294 (2008).
19. WallisDeVries, M. F. & van Swaay, C. A. M. Global warming and excess nitrogen may induce butterfly decline by microclimatic cooling. *Glob. Change Biol.* **12**, 1620–1626 (2006).
20. Warren, M. S. & Bourn, N. A. D. Ten challenges for 2010 and beyond to conserve Lepidoptera in Europe. *J. Insect Conserv.* **15**, 321–326 (2011).
21. Thomas, J. A., Simcox, D. J. & Clarke, R. T. Successful conservation of a threatened *Maculinea* butterfly. *Science* **325**, 80–83 (2009).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

22. Bennett, N. L., Severns, P. M., Parmesan, C. & Singer, M. C. Geographic mosaics of phenology, host preference, adult size and microhabitat choice predict butterfly resilience to climate warming. *Oikos* **124**, 41–53 (2015).
23. McBride, C. S. & Singer, M. C. Field studies reveal strong postmating isolation between ecologically divergent butterfly populations. *PLoS Biol.* **8**, e1000529 (2010).
24. Harrison, S. Long-distance dispersal and colonization in the Bay Checkerspot butterfly, *Euphydryas editha bayensis*. *Ecology* **70**, 1236–1243 (1989).
25. Thomas, C. D., Singer, M. C. & Boughton, D. A. Catastrophic extinction of population sources in a butterfly metapopulation. *Am. Nat.* **148**, 957–975 (1996).
26. Nylin, S., Slove, J. & Janz, N. Host plant utilization, host range oscillations and diversification in nymphalid butterflies: a phylogenetic investigation. *Evolution* **68**, 105–124 (2014).
27. Gonzalez, A., Ronce, O., Ferriere, R. & Hochberg, M. E. Evolutionary rescue: an emerging focus at the intersection between ecology and evolution. *Phil. Trans. R. Soc. Lond. B* **368**, 20120404 (2013).
28. Carroll, S. P. et al. Applying evolutionary biology to address global challenges. *Science* **346**, 1245993 (2014).
29. Singer, M. C. Butterfly–host relationships: host quality, adult choice and larval success. In *The Biology of Butterflies, Symp. Roy. Ent. Soc. 11* (eds Vane-Wright, R. I. & Ackery, P. R.) 81–88 (Academic, London, 1984).
30. Newcombe, R. G. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat. Med.* **17**, 857–872 (1998).

Acknowledgements P. R. Ehrlich introduced M.C.S. to *E. editha* in 1967; C. D. Thomas, H. L. Billington, D. Ng, L. E. Gilbert and J. L. B. Mallet helped to initiate the project. D. D. Murphy performed the 2010 census. C.L. Boggs, R. A. Steward, J. L. B. Mallet and C. S. McBride critiqued the manuscript. The 2013 Integrative Biology Faculty Merit Review Committee, University of Texas at Austin, provided incentive to complete the study. J. Schneider and B. Schneider, the Drudge family, Clear Creek Tahoe and The Nature Conservancy allowed access to the site.

Reviewer information *Nature* thanks T. Oliver, M. Friberg and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Both authors participated in field censuses and writing. M.C.S. performed oviposition preference tests and statistical analyses.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0074-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0074-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.C.S. or C.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Statistics. No statistical methods were used to predetermine sample size. All tests are two-tailed. Error bars on figures—given only where justified by sampling techniques—are 95% confidence interval (c.i.), calculated using http://vassarstats.net/ and the methodology of Newcombe[30] for c.i. of proportions, with continuity correction. Figure 1a does not explicitly depict insects without preference but they can be deduced by subtracting the depicted percentages from 100%. Otherwise no relevant data have been omitted from any experiment or set of observations. Sample sizes were largely limited by feasibility.

Blind preference testing. Oviposition preference tests to estimate heritability[17] were performed blind, in the sense that the tester did not know which insects were siblings or offspring of particular parents. The high heritability estimate from this blind testing (0.9) gave confidence that the tests were not subject to severe observer bias; repeated blind testing of the same butterfly by different observers has also given confidence. However, in the present analyses the appropriate blind test would require the tester to not know which year it was, and we were not able to achieve this without compromising the quality of the data.

Oviposition preferences. To generate the data shown in Fig. 1a, butterflies were captured in the field and their oviposition preferences tested by a standardized technique, in which encounters are staged between the tested insect and each plant in alternation. Plants were undisturbed in their natural habitats or freshly transplanted into pots in their own soil. Acceptance of plant taste was judged from full abdominal curling and extrusion of the ovipositor for 3 s[31]. Acceptance and rejection were recorded at each encounter, but oviposition was not allowed[31]. Videos showing acceptance in such staged encounters have previously been published[23]. During each test the range of plants that would be accepted, if encountered, expands over time with increasing motivation to oviposit. Therefore, acceptance of plant A followed by rejection of plant B is recorded as preference for A over B. Testing of assumptions underlying this technique have previously been described[31]. Because insects without preference are not shown in Fig. 1a, percentages do not sum to 100%, except for 2005 and 2007, when preference for *Plantago* was unanimous among tested butterflies. Raw data are shown in Extended Data Table 3. A more detailed comparison between early and late periods, showing strength as well as direction of preferences, is given in Extended Data Fig. 7. The assumption that the oviposition preferences of these insects are not influenced by prior experience, either as larvae or as adults, is supported by previous observations and experiments[17,31,32].

Percentages of egg clutches and/or larval groups found on *Plantago*. Raw data for Fig. 1b are shown in Extended Data Tables 1, 3. Low sample sizes in 1988 and 1989 were total counts, reflecting a population bottleneck. The high proportion of larvae on *Plantago* in 1989 reflects a lag in recolonization of *Collinsia* after the spatial restriction in the bottleneck, shown in Extended Data Fig. 2. In 2002–2007 we searched a larger area for larvae on *Collinsia* than for larvae on *Plantago*. For these years, Fig. 1b—showing the proportion of larvae found on the two hosts—overestimates the overall proportion on *Collinsia*; the areas searched are given in Extended Data Table 1. To be conservative, we indicated data for 2002 and 2005 as 'published' although neither sample sizes nor confidence limits had previously been given[32]. The graph shows that in 2015 we found a single group of hatching eggs on *Collinsia*. However, we performed no census in 2015 as our visit was too early.

Estimates of density of larval webs. Figure 1c gives estimates of the density of larval webs on all hosts combined per 10,000 m^2. Raw data are given in Extended Data Table 1. In each census, individual host plants of all species were searched for eggs, larvae, larval webs and typical damage. Very different scales of census were conducted in different years. For small patches, every individual host plant could be searched; for larger patches, stratified line transects were used. Confidence limits are not given, but the logarithmic scale of the *y* axis makes clear the scale of population changes; for example, the raw data (Extended Data Table 1) include the finding of 34 groups in only 70 m^2 in 1982 and 4 groups in 20,000 m^2 in 1988. No density estimate is given for 2007 because, although 38 groups were found (Fig. 1b), at the time of the last census some were still eggs, and so were harder to find than larvae. Sample calculation for 2002 (see Methods for Extended Data Table 1): estimated number of webs on *Plantago* = 67 × 4,000/170 = 1,576. Total number on *Collinsia* = 3 (all were counted). Estimated total number webs in entire area of 20,000 m^2 = 1,579; density per 10,000 m^2 = 789.

Survival of *E. editha* larvae on *Collinsia* and *Plantago*. Effects of oviposition host on fitness are given in Table 1. Eggs were placed out on randomly chosen hosts in the field, by manipulating butterflies to lay (see previously published videos[23]). Gregarious neonate larvae were placed with a sable brush in groups of 20–35. After ten days of larval life (just before diapause), each group was gathered. It is in pre-diapause life that the principal effects of oviposition on fitness are manifest, as post-diapause larvae are mobile and can switch between host species.

Estimated host densities and percentages of embedded and exposed hosts. Host densities were estimated from quadrats along line transects placed within the strata of the ecotone in which each plant was concentrated (Extended Data Fig. 1). Early approximate estimates of the percentage of embedded plants were derived from photographs, memory and anecdotal observation, because embeddedness was previously uncommon and its value as a trait was not anticipated before the sudden embedding of *Plantago* in 2007. A plant was classed as embedded if it was surrounded for >50% of its circumference by vegetation taller than itself. In most cases the differences between embedded and exposed plants were striking (Extended Data Fig. 3); in the March census in 2007, some embedded plants were hidden and not found—they were pressed under thatch from winter snow and inaccessible to post-diapause larvae feeding in March—but reappeared and, if they were no longer completely embedded, were accessible to ovipositing butterflies in May. Extended Data Fig. 4a has a photo of eggs naturally laid on such a reappearing *Plantago*, emerging from winter thatch, in May, 2007.

Numbers of groups of *E. editha* on *Plantago*, *Collinsia* and *Penstemon*. Extended Data Table 1 contains counts over wider areas and including lower host densities than the 'core' areas censused in Table 2. The total area within which suitable *Plantago* might be found was around 4,000 m^2; the total area that might contain suitable *Collinsia* was larger, maximally about 17,000 m^2 (1982 map in Extended Data Fig. 2) but less in dry years. About 1,000 m^2 overlapped between the two distributions. Wider areas were searched in 1988 and 1989 to check whether we had missed part of the population in previous work. We had not, so as we found no habitat in this wider search, subsequent searches were restricted to meadow margins and adjacent sagebrush, approximately 20,000 m^2 (Extended Data Figs. 1, 2).

In 2002, 2005 and 2007, the entire area in which larvae might have been found on *Collinsia* was searched; *Plantago* areas were not searched in their entirety, but merely sufficiently to get an estimate of plant density and occupancy by the butterflies. After the extinction in 2007–2008, the entire habitat was searched in each census. In most years, more plants were searched for *E. editha* than were included in censuses to estimate plant density reported in Table 2.

Extended Data Table 1 shows data from areas censused in which plants with and without larvae were counted. Where maps in Extended Data Fig. 2 show more insects than in Extended Data Table 1, as in 1982, the insect distribution in the map is derived from rapid assessments in which insects were observed and counted but plants were not.

Temperatures of eggspaces on *Collinsia* and *Plantago*. Eggspace temperatures given in Extended Data Table 2 were measured with a fine thermocouple (MT-29/1B insect probe, type T, copper-constantan). After measuring each exposed *Plantago* we measured one or two adjacent (within 1 m) embedded plants growing in the same position within the ecotone; we took care that embedded plants were not systematically growing in more humid microsites. When two embedded plants were measured, we used the mean value of the two for analysis.

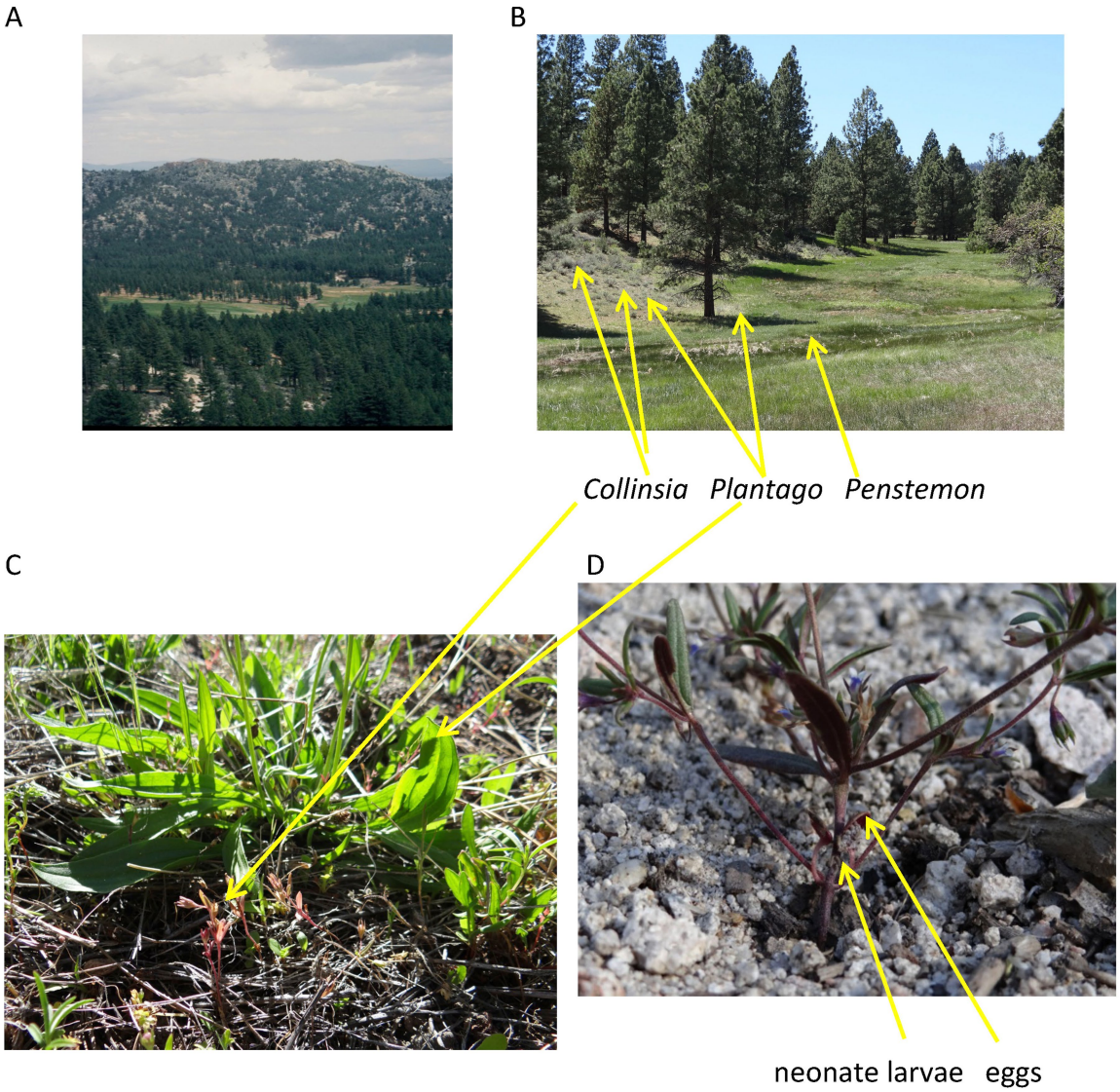
We found three errors in our previous publications: (1) number of groups found on *Collinsia* in 1990 is here corrected to 18 from 6[6]; (2) number of butterflies preference-tested in 1986 is corrected to 36 from 31[5,6]; (3) a misleading early reference to the population as feeding on *Plantago* in 1969[33] stemmed from initial visit in June, after *Collinsia* had senesced and disappeared. Faster-growing larvae on *Collinsia* had entered diapause and were not found, although some still remained feeding on *Plantago*. The next visit in 1971 made it clear that although *Plantago* was already being used, *Collinsia* was the principal host. No censuses were performed in 1969–1971.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Raw data are included in Fig. 1, Tables 1, 2, Extended Data Figs. 1–7 and Extended Data Tables 1–3. All other data are available from the corresponding authors upon reasonable request.

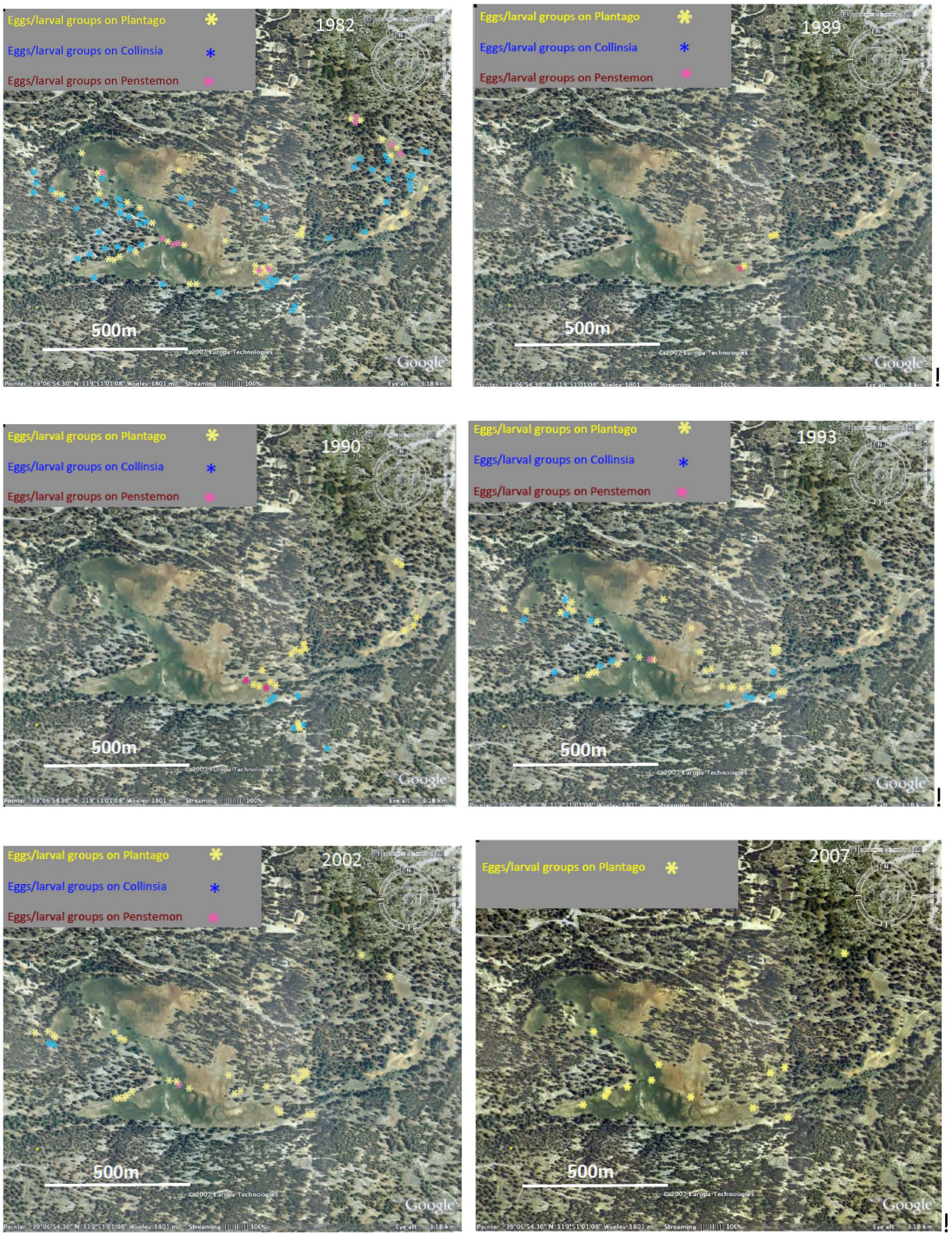
31. Singer, M. C., Vasco, D. A., Parmesan, C., Thomas, C. D. & Ng, D. Distinguishing between preference and motivation in food choice: an example from insect oviposition. *Anim. Behav.* **44**, 463–471 (1992).
32. Singer, M. C., Wee, B., Hawkins, S. & Butcher, M. in *The Evolutionary Ecology of Herbivorous Insects: Speciation, Specialization and Radiation* (ed. Tilmon, K. J.) 311–324 (Univ. of California Press, Berkeley, 2008).
33. Singer, M. C. Evolution of food-plant preference in the butterfly *Euphydryas editha*. *Evolution* **25**, 383–389 (1971).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



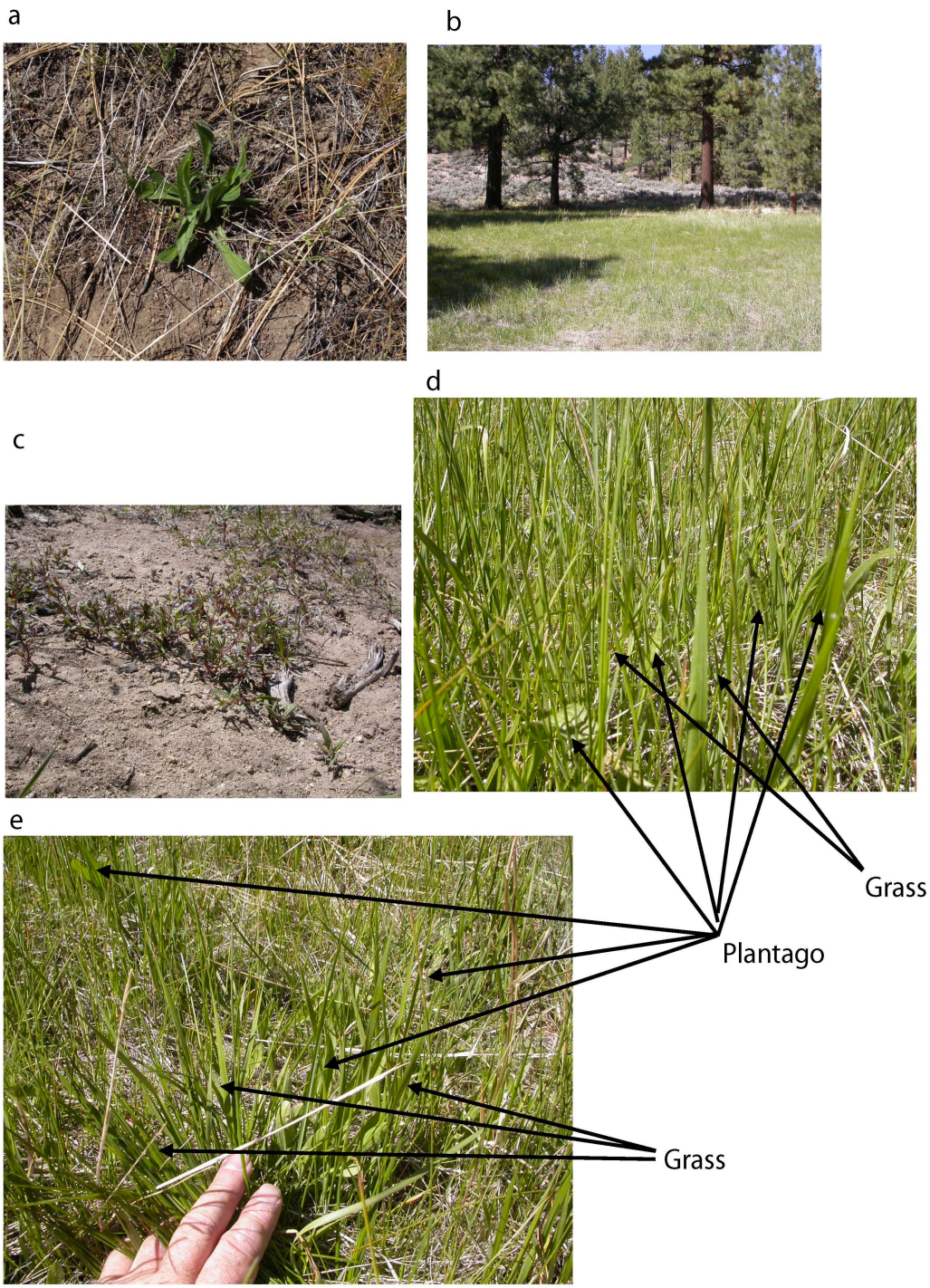
Extended Data Fig. 1 | Habitat and host distributions at Schneider's Meadow. a, The butterfly habitat is a single, isolated, spring-fed wet meadow in the centre of the photograph, surrounded by non-habitat for the butterflies: dry sagebrush scrub and coniferous forest. **b**, Distribution in the meadow-edge ecotone of the principal hosts, *Plantago* and *Collinsia*, plus the minor host *Penstemon*. **c**, Typical difference in phenology between *Plantago* and *Collinsia* in May 2014. In the foreground are red, senescent *Collinsia* plants that are edible to the insects but which will die within a few days; behind them is a single green, budding *Plantago* that will remain edible until after all *E. editha* larvae have entered diapause. **d**, Hatching egg clutch on *Collinsia* cotyledon in hot, dry microhabitat.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



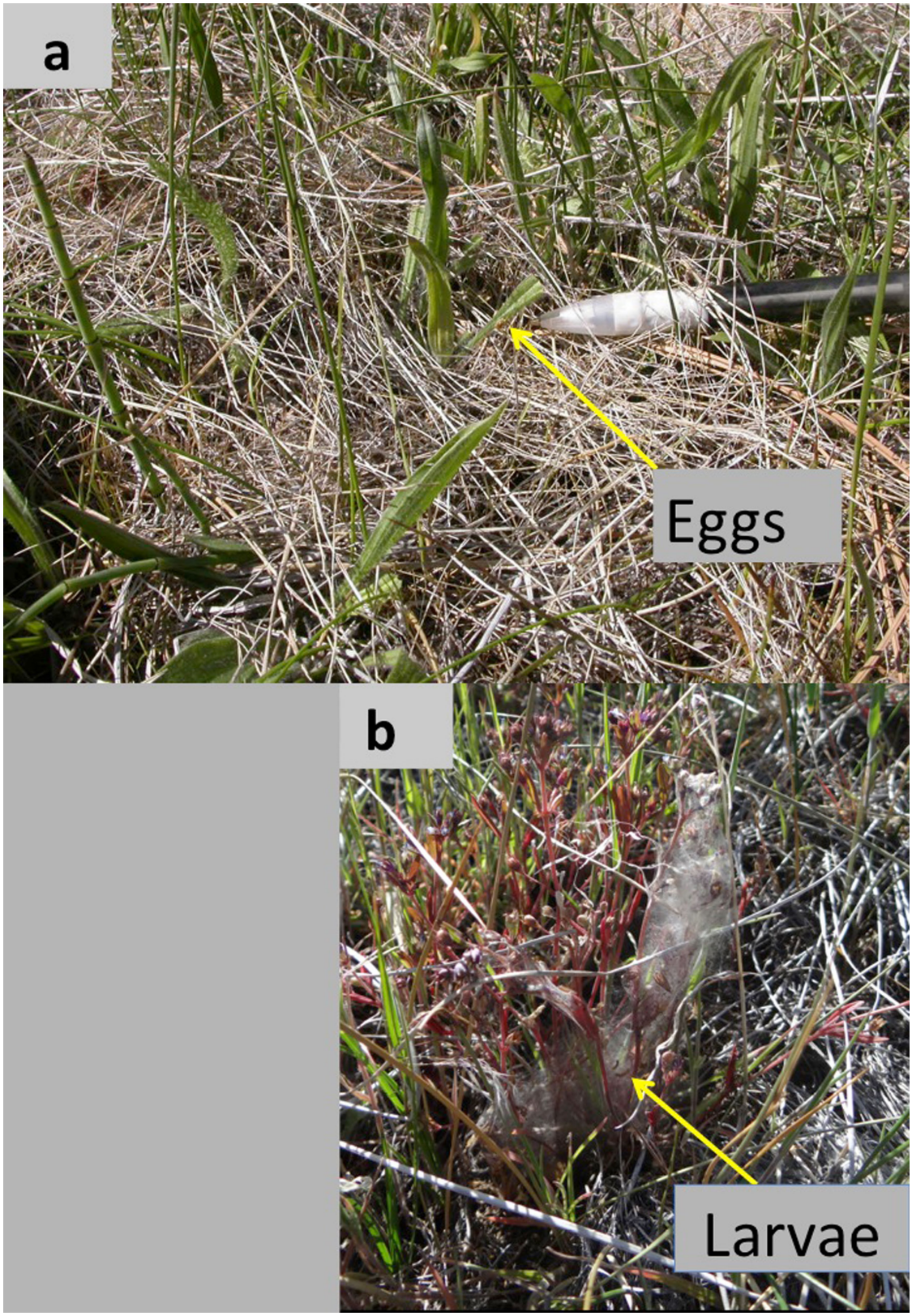
Extended Data Fig. 2 | Changes in distribution of early stages of *E. editha* (eggs or larvae) from 1982 to 2007. Data were added by hand to the GoogleMaps image. Most stars represent several groups. For example, in 1989, 23 groups were found on *Plantago* and one on *Penstemon*. The restricted distribution in that year followed a bottleneck in 1988 after record-breaking cold in January, without the usual insulating snow cover. Schneider's Meadow is at 1,700 m elevation: nearby towns at lower elevations recorded −25 °C on 1 January 1988 (Minden, 1,444 m elevation) and −20 °C on 18 January (Carson City, 1,424 m). Note the recolonization of *Collinsia* as the insects expanded back into the distribution of *Collinsia* in 1990 and 1993. Larval groups recorded in 1988 and 1989 were clustered around an attractive nectar source (*Wyethia* sp.); it is possible that adults attracted to this nectar in 1988 had survived as larvae on *Collinsia* in 1987–1988 and then, as adults, laid eggs in 1988 on *Plantago* adjacent to nectar. This possibility prevents us from making a definite conclusion that the population would have become extinct if eggs laid in 1987—before the bottleneck—had been placed only on *Collinsia*. Data for 2005 exist and closely resemble those for 2002.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 3 | Effects of cessation of grazing: *Plantago* plants embedded, whereas *Collinsia* plants are unaffected. Data are provided in Table 2. **a**, *Plantago* at Schneider in 1984, exposed to full sunlight and physically acceptable to ovipositing *E. editha*. **b**, Meadow edge in May 2007, after cattle removal. In the foreground is *Plantago* habitat with thick grasses; in the background is *Collinsia* habitat not grassed-in, with barren spaces between the sagebrush. **c**, *Collinsia* in May 2007, unaffected by the embedding that simultaneously affected the *Plantago* plants shown in **d** and **e**. Embedding in grasses not only cooled the *Plantago* plants (Extended Data Table 2) but also rendered them hard to find, both by butterflies seeking oviposition sites and by larvae seeking food.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 4 | Egg clutch and communal web. a, Natural egg clutch laid in May 2007 on *Plantago*. The plant is pushing through winter thatch, and would have been unlikely to be acceptable to ovipositing butterflies before cattle removal, when plants similar to the one in Extended Data Fig. 3a were available. **b**, Communal web spun after recolonization. Second-instar larvae on *Collinsia* at Schneider in May 2014. This is a single group of larvae, probably stemming from a single oviposition event; there were nine such groups, all on *Collinsia*. Unexpectedly, this group is not on the most exposed *Collinsia* available.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



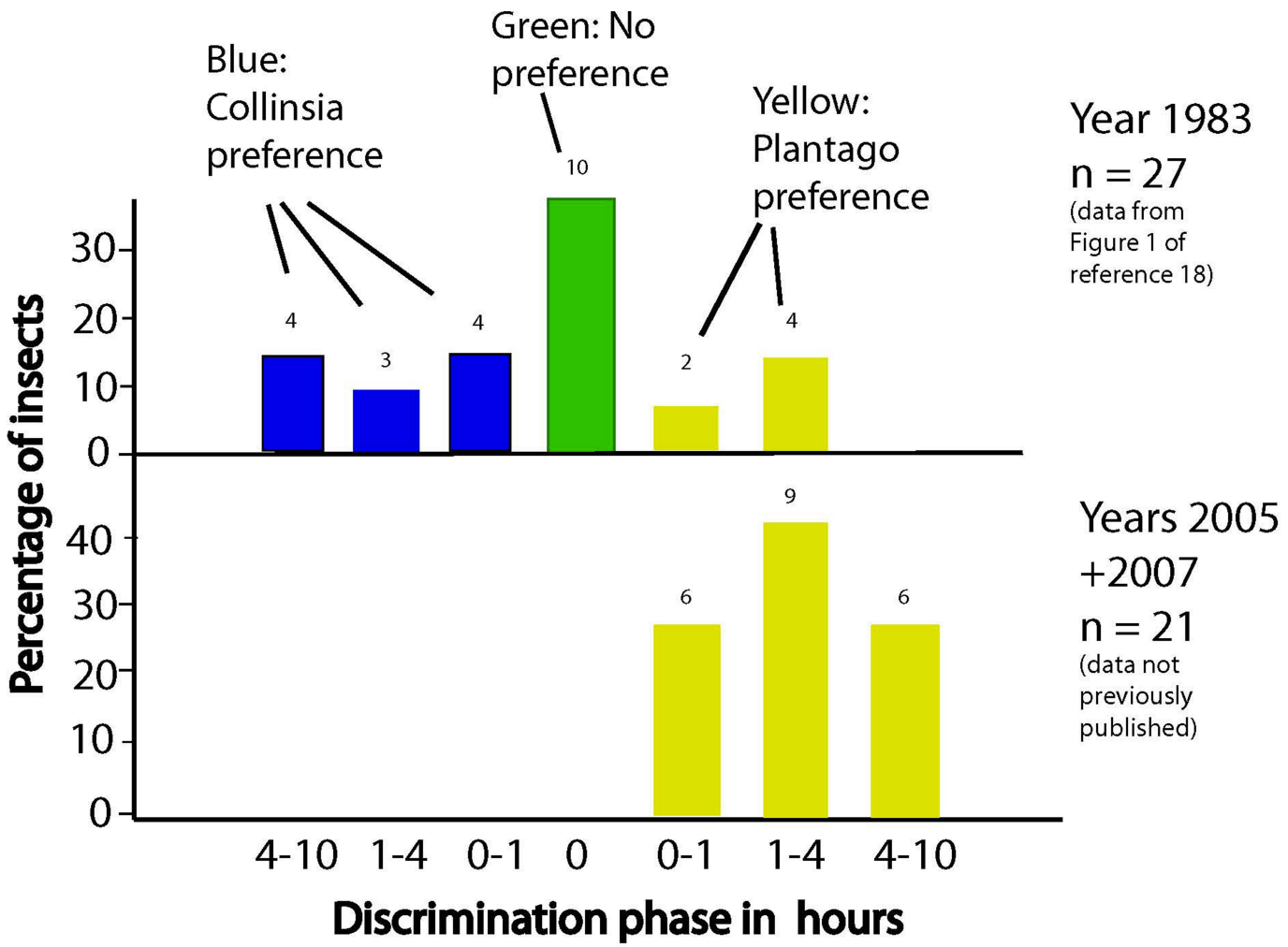
Extended Data Fig. 5 | The return of mostly exposed *Plantago* plants after anthropogenic lushness abated. Photographs were taken in 2014, but Table 2 shows that they could have been taken in 2008 or subsequently.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Effect of fire on size and longevity of *C. parviflora*. A single fire-enhanced *Collinsia* at McGee Creek (east of Bishop, California) is still blooming. There is a small web of *E. editha* larvae at its base from a naturally laid egg clutch. The fifteen senescent individual *Collinsia* lying on the ground represent a haphazard sample gathered from unburned microsites within 2 m of the enhanced individual.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Strength and direction of the oviposition preferences of butterflies sampled at Schneider in 1983 and 2005 + 2007. The number over each bar is the sample size of biologically independent samples: individual butterflies captured in the field. The discrimination phase is the length of time for which the insect would search, during which it would consistently accept the preferred host and consistently reject the second-ranked host. At the end of this phase, if it does not succeed in ovipositing, the insect enters an acceptance phase after reaching the level of oviposition motivation at which either host would be accepted (whichever was next encountered). Insects in the blue 1–4 column would search for 1–4 h, during which only *Collinsia* would be accepted. If they failed to find *Collinsia* within 4 h, they would subsequently accept either host, until actual oviposition occurred. Green central bar shows butterflies without preference. Sample size for 2005–2007 is smaller than in Fig. 1a because we include on Fig. 1a (and here omit) five butterflies for which we determined the direction of preference, but not the strength.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Census results: areas searched and numbers of egg clutches or larval webs found on each host

Year (date)	Approximate area searched (m^2)	Groups on *Plantago* *(approximate number of plants searched)*	Groups on *Collinsia* *(approximate number of plants searched)*	Groups on *Penstemon* *(approximate number of plants searched)*
1982 (19-21 May)	70	4 (67)	28 (441)	2 (233)
1983 (2 June)	20	3 (89)	11 (506)	0 (112)
1984 (17-31 May)	1,250	4 (79)	27 (1169)	0 (94)
1986 (26 April-17 May)	Not recorded	62 (not counted)	50 (not counted)	10 (not counted)
1987 (7-8 May)	2,030	8 (260)	10 (892)	0 (190)
1988	20,000	3 (434)	0 (>3,000)	1 (87)
1989	50,000	23 (462)	0 (5,500)	1 (504)
1990 (13 May)	2,000	25 (135)	18 (1789)	2 (365)
1993 (1-3 June)	5,800	22 (1104)	9 (1390)	1 (116)
New data below here				
2002	170 for *Plantago* 16,000 for *Collinsia*	64 (770)	3 (>5000)	1 (64)
2005 (6-12 June)	65 for *Plantago* 16,000 for *Collinsia*	67 (317)	1 (>3000)	0 (209)
2007 (17-18 March, 5 May)	800 for *Plantago,* 16,000 for *Collinsia*	38 (211)	0 (>2000)	0 (30)
2008	4,000 for *Plantago,* 16,000 for *Collinsia*	0 (2115)	0 (>4,000)	0 (88)
2009	4,000 for *Plantago,* 16,000 for *Collinsia*	0 (1840)	0 (>3,000)	0 (31)
2014 (11-12 May)	4,000 for *Plantago,* 16,000 for *Collinsia*	0 (1125)	9 (>3,000)	0 (46)
2015 (28-31 May)	n/a	0 (n/a)	1 (n/a)	0 (n/a)

Data for 1982–1993 have previously been published[6]. These are not the same data shown in Extended Data Fig. 2 (see Methods). Extension of survey area to 50,000 m^2 in 1989 did not reveal additional habitat, so density in Fig. 1c for that year is calculated using the estimated maximum habitat area of 20,000 m^2, giving a higher estimate than that previously reported[6].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Measurements of temperature excess over ambient air at 1-m height at three potential types of oviposition site at Schneider's Meadow: embedded *Plantago*, exposed *Plantago* and exposed *Collinsia*

	native *Collinsia*	exotic *Plantago*
exposed	15.6, 10.8, 9.5, 14.8, 8.8, 8.4, 10.2, 12.1, 10.4, 12.8, 12.9,11.2. N = 12	11.6,11.1,11.0,12.2,9.4, 15.5,17.2,13.1,9.2,10.8, 11.8,17.4,11.2,19.3,16.5, 13.9,19.0,10.6 N = 18
embedded		7.9,5.2,7.2, 8.3, 4.1, 9.4, 5.1, 9.7, 5.3, 1.2, 2.5, 6.0, 3.1, 5.4, 7.9, 9.2, 3.3, 7.5 N = 18

Measurements were taken at eggspace height (2–4 cm) and at 1 m above ground between 12:00 and 15:00 in May 2015. Each measurement of exposed *Collinsia* or *Plantago* came from a different, haphazardly chosen habitat patch. On occasion we measured two embedded *Plantago* plants in the same patch; when this occurred we show the mean of the two values and used these means in analysis, treating the two plants as a single sample. Consequently, each data point represents a biologically independent sample, and the independent sample sizes are 12, 18 and 18 as indicated at the base of each cell.

Comparison between exposed *Plantago* (mean excess temperature above ambient = 13.4 °C) and embedded *Plantago* (mean excess temperature above ambient = 6.0 °C) by two-sided *t*-test: $t = 7.55$, degrees of freedom = 28, $P < 0.0001$.

Comparison between exposed *Collinsia* (mean excess temperature above ambient = 11.46 °C) and exposed *Plantago* (13.4 °C) by two-sided *t*-test: $t = 1.77$, degrees of freedom = 28, $P = 0.09$.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 3 | Source data for means and confidence limits of adult preference and larval diet

Year	Sample size adult females	Number preferring Plantago	Proportion preferring Plantago	Lower 95% c.i.	Upper 95% c.i.
1982	44	3	0.068	0.0178	0.197
1983	34	8	0.235	0.114	0.416
1985	59	12	0.203	0.114	0.332
1986	36	11	0.306	0.169	0.483
1990	27	14	0.519	0.324	0.708
2005	11	11	1.00	0.678	1.00
2007	15	15	1.00	0.747	1.00
Year	Sample size adult females	Number preferring Collinsia	Proportion preferring Collinsia	Lower 95% c.i.	Upper 95% c.i.
1982	44	27	0.614	0.455	0.752
1983	34	16	0.471	0.301	0.646
1985	59	25	0.423	0.299	0.559
1986	36	16	0.444	0.283	0.617
1990	27	5	0.185	0.070	0.388
2005	11	0	0	0	0.321
2007	15	0	0	0	0.254
Year	Sample size larval groups	Number of larval groups on Plantago	Proportion of groups on Plantago	Lower 95% c.i.	Upper 95% c.i.
1984	31	4	0.129	0.042	0.308
1986	122	62	0.508	0.417	0.599
1987	18	8	0.444	0.224	0.687
1990	45	25	0.556	0.401	0.700
1993	32	22	0.688	0.499	0.833
2002	69	64	0.928	0.832	0.973
2005	68	67	0.985	0.910	0.999
2014	9	0	0	n/a (complete census)	

Source data for Fig. 1a, b. Statistical test described in the legend of Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0084-4

Subepithelial telocytes are an important source of Wnts that supports intestinal crypts

Michal Shoshkes-Carmel[1], Yue J. Wang[1], Kirk J. Wangensteen[1], Beáta Tóth[2], Ayano Kondo[1], Efi E. Massassa[2], Shalev Itzkovitz[2] & Klaus H. Kaestner[1]*

Tissues that undergo rapid cellular turnover, such as the mammalian haematopoietic system or the intestinal epithelium, are dependent on stem and progenitor cells that proliferate to provide differentiated cells to maintain organismal health. Stem and progenitor cells, in turn, are thought to rely on signals and growth factors provided by local niche cells to support their function and self-renewal. Several cell types have been hypothesized to provide the signals required for the proliferation and differentiation of the intestinal stem cells in intestinal crypts[1–6]. Here we identify subepithelial telocytes as an important source of Wnt proteins, without which intestinal stem cells cannot proliferate and support epithelial renewal. Telocytes are large but rare mesenchymal cells that are marked by expression of FOXL1 and form a subepithelial plexus that extends from the stomach to the colon. While supporting the entire epithelium, FOXL1$^+$ telocytes compartmentalize the production of Wnt ligands and inhibitors to enable localized pathway activation. Conditional genetic ablation of porcupine (*Porcn*), which is required for functional maturation of all Wnt proteins, in mouse FOXL1$^+$ telocytes causes rapid cessation of Wnt signalling to intestinal crypts, followed by loss of proliferation of stem and transit amplifying cells and impaired epithelial renewal. Thus, FOXL1$^+$ telocytes are an important source of niche signals to intestinal stem cells.

After the cloning of *Foxl1*, which encodes the gene for the winged-helix transcription factor FOXL1 (previously known as Fkh6[7]), initial expression profiling focused on measuring its steady-state mRNA levels, both by RNase protection assays and by in situ hybridization[8,9]. *Foxl1* transcripts were found at low levels in the adult stomach and intestine, and in about one to two cell layers of the mesodermal tissue surrounding the primitive fetal gut tube. In order to visualize subepithelial *Foxl1*-expressing cells in their entirety, we used the *Foxl1-cre* transgenic line[10] in conjunction with genetic lineage labelling using the *Rosa26-mTmG* allele[11]. In this model, Cre-labelled cells produce a membrane-targeted version of green fluorescent protein (GFP), which allows us to map the size and location of *Foxl1*-expressing cells (Fig. 1a). Figure 1b shows an example of duodenal crypts, where *Foxl1*-expressing cells, identified both by nuclear FOXL1 staining (white) and membraneous GFP expression (green), are in close contact with the entire crypt base, identified by EPCAM staining (red). FOXL1$^+$ cells are large, flat cells with cytoplasmic processes in excess of 100 μm. Cells with these properties, that is, very large mesenchymal cells with extended cell bodies, were identified previously by electron microscopy and termed 'telocytes', with the long processes termed 'telopodes'[12,13].

Intestinal telocytes express platelet-derived growth factor receptor alpha (PDGFRα)[14]. PDGFRα staining overlapped with *Foxl1-cre*-induced GFP expression, further supporting the notion that these cells are intestinal telocytes (Fig. 1c, d), although PDGFRα is also present in numerous cells deeper in the submucosa that do not express *Foxl1*. Immunoelectron microscopy confirmed that FOXL1$^+$ cells form long telopodes that encompass all crypt cells, and are separated from the epithelium by sub-micrometre distances (Fig. 1e–g).

In order to gain a better understanding of the three-dimensional nature of the FOXL1$^+$ telocyte network, we performed confocal imaging of immunofluorescence-stained jejunum following tissue clearing. FOXL1$^+$ PDGFRα$^+$ telocytes form a plexus that supports the entire epithelium, visualized by EPCAM staining (Fig. 1h and Supplementary Video). Thus, FOXL1$^+$ PDGFRα$^+$ positive telocytes maintain a subepithelial sheath that is juxtaposed to the epithelium, and therefore in an ideal position to act as a niche to provide signals to the stem and progenitor cells in the intestinal crypt.

For the molecular characterization of FOXL1$^+$ telocytes, we sorted cells based on *Foxl1-cre;Rosa26-mTmG* activity (Extended Data Fig. 1), performed RNA sequencing analysis (RNA-seq), and compared the expression profiles to those of FOXL1$^-$ mesenchymal cells, LGR5$^+$ stem cells (sorted by *Lgr5-eGFP^{High15}* expression) and epithelial enterocytes (Fig. 2a). Note that while FOXL1$^+$ cells express high levels of *Pdgfrα* and *Cd34*, these two genes are also active in FOXL1$^-$ cells, confirming our observation from immunostaining and genetic lineage tracing that FOXL1$^+$ cells constitute a small subset of PDGFRα$^+$ cells. FOXL1$^+$ telocytes express high levels of multiple regulators—both activators and repressors—of key signalling pathways, such as Wnt, SHH, Bmp and TGF-β (Fig. 2b–g). While the Wnt ligands WNT2B and WNT5A are specifically expressed in FOXL1$^+$ telocytes, the Wnt signalling enhancer and LGR5 ligand R-spondin3 is produced by both FOXL1$^+$ telocytes and FOXL1$^-$ mesenchymal cells (Fig. 2c; Extended Data Table 1). Notably, FOXL1$^+$ telocytes also express the Wnt inhibitors DKK3 and SFRP1. Similarly, FOXL1$^+$ telocytes express BMP4, BMP5, BMP6 and BMP7 as well as the Bmp inhibitors chordin-like1 and gremlin1 (Fig. 2d).

The results of the expression profiling of FOXL1$^+$ telocytes summarized above present an apparent paradox. How can FOXL1$^+$ telocytes provide Wnt ligands to activate Wnt signalling in intestinal crypts if they also produce Wnt pathway inhibitors? We speculated that there might be a regional differentiation of the transcriptome of FOXL1$^+$ telocytes according to their position along the crypt–villus axis. In order to test this hypothesis, we performed single-molecule fluorescence in situ hybridization (smFISH) to detect mRNA encoding activators (WNT2B, WNT5A and R-spondin 3) and inhibitors (SFRP1 and DKK3) of the Wnt signalling pathway, and BMP5, which signals via activation of the Smad signalling cascade. We labelled telocytes by immunostaining for PDGFRα, as FOXL1 is a nuclear protein and thus does not show the full extent of the cells. As shown in Fig. 3a, telocytes express both *Wnt2b*, an activator of canonical Wnt signalling, and *Wnt5a*, which acts via the planar cell polarity pathway. However, the localization of the two mRNAs is clearly distinct, with *Wnt2b* enriched at the bottom of the crypts where LGR5$^+$ stem cells are situated, and *Wnt5a* mRNA present near the crypt–villus junction. Likewise, there is a striking pattern of co-expression of *Wnt2b* with *Sfrp1*, which encodes a secreted frizzled protein that acts as a decoy receptor to limit Wnt signalling at the bottom of the crypt; whereas towards the crypt–villus junction, *Sfrp1* mRNA predominates, as the density of *Wnt2b* mRNA

[1]Department of Genetics and Center for Molecular Studies in Digestive and Liver Diseases, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. *e-mail: kaestner@pennmedicine.upenn.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

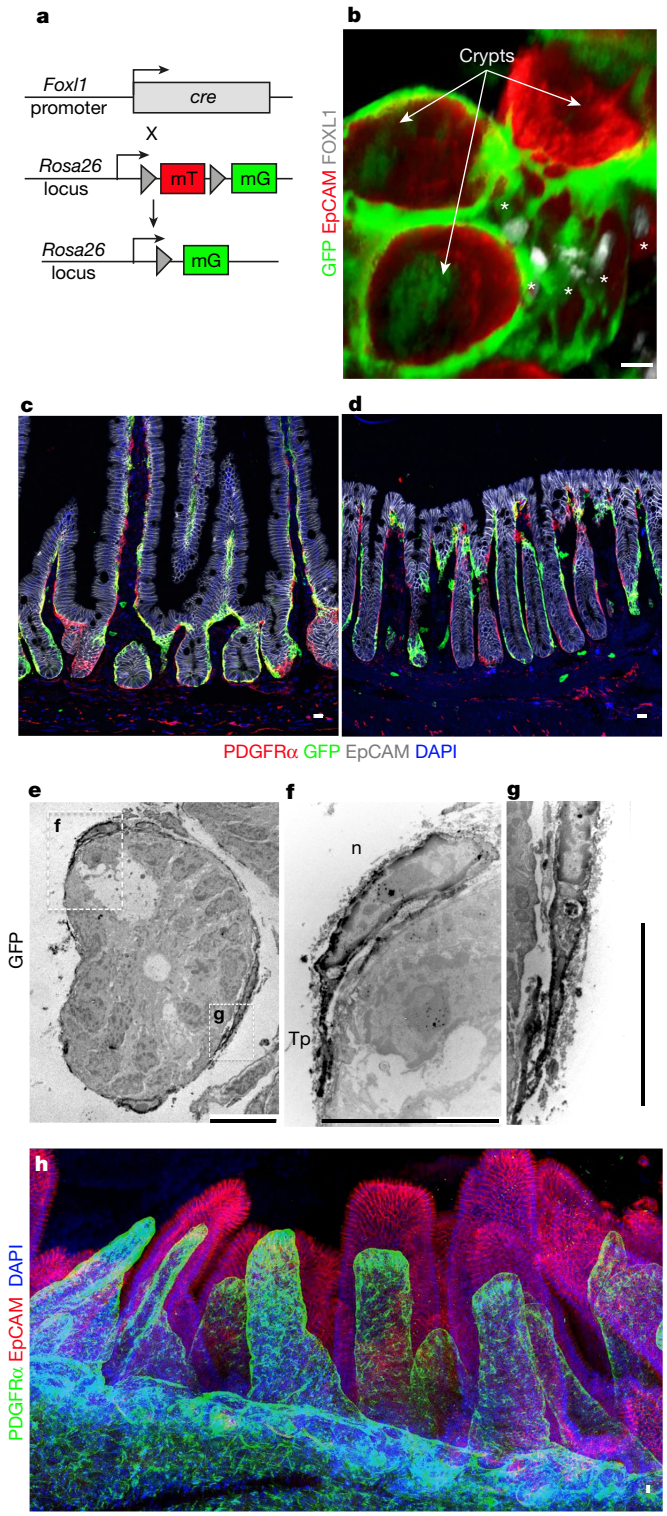


Fig. 1 | FOXL1$^+$ cells are telocytes and co-express PDGFRα. a, Schema for labelling FOXL1$^+$ cells with GFP. **b**, *Foxl1-cre*-driven GFP expression is restricted to pericryptal mesenchymal telocytes. FOXL1 (white, *), GFP (green) and EPCAM (red) immunofluorescence in cleared whole mouse duodenum. **c**, **d**, GFP (green) and PDGFRα (red) immunofluorescence in the duodenum (**c**) and colon (**d**). **e**–**g**, Immunoelectron microscopy showing GFP in duodenal crypt (transverse section) of a *Foxl1-cre;Rosa26-mTmG* mouse. **f**, Inset showing the nucleus and the telopode of a telocyte. n, nucleus; Tp, telopode. **g**, Inset showing contact between two telopodes. **h**, Confocal imaging of cleared whole small intestine showing expression of PDGFRα (green) and EPCAM (red). Experiments were repeated at least three times with similar results. Scale bars, 10 μm.

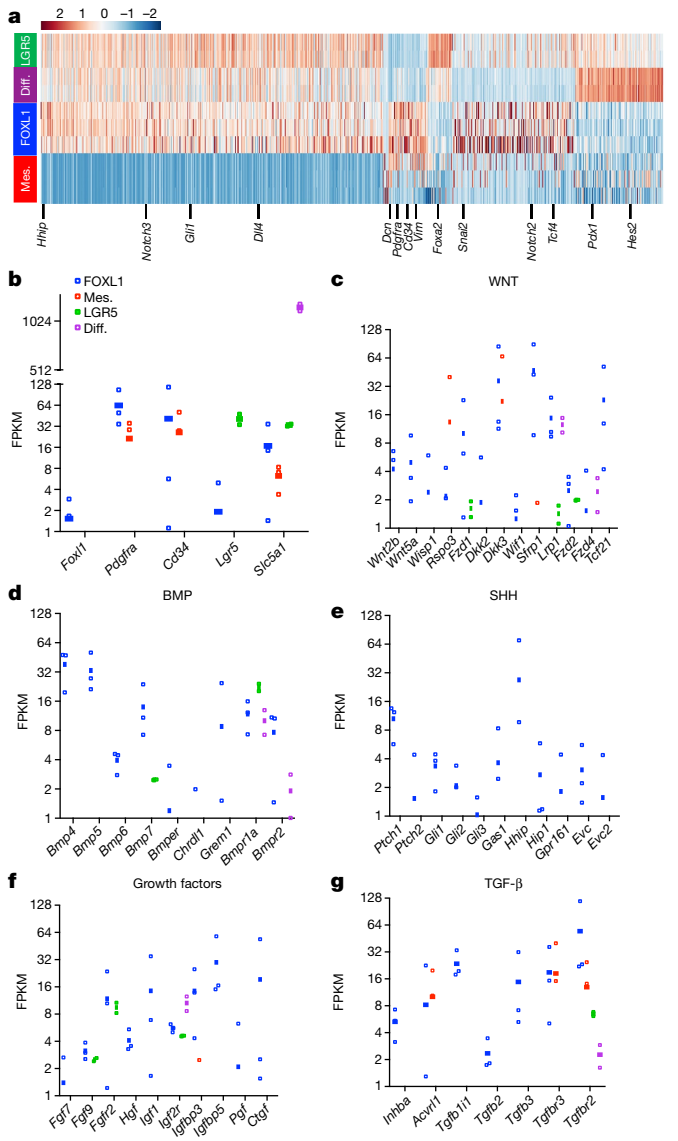


Fig. 2 | Telocytes express key signalling molecules. RNA-seq analysis of FOXL1$^+$ cells ($n = 3$ mice) compared to FOXL1$^-$ mesenchymal cells (Mes., $n = 3$ mice), LGR5$^+$ stem cells ($n = 2$ mice) and differentiated enterocytes (Diff., $n = 2$ mice). **a**, Heat map showing hierarchical clustering. Only genes with more than tenfold difference in expression level in pairwise comparisons and a false discovery rate (FDR) of less than 2% were used for the analysis. Values are shown as normalized *Z* score ± s.d. **b**–**g**, Selected markers showing differential expression: *Foxl1*, *Pdgfra* and *Cd34* (telocyte markers), *Lgr5* and *Slc5a1* (**b**), WNT pathway (**c**), BMP pathway (**d**), SHH pathway (**e**), growth factors (**f**) and TGF-β pathway (**g**). Data are shown as fragments per kilobase million (FPKM). Error bars indicate s.e.m.

molecules decreases (Fig. 3f, j). *Rspo3*, which encodes an important co-activator of the Wnt pathway, is present in telocytes only at the bottom of the crypt, and is absent further up the crypt–villus axis (Fig. 3c). By contrast, transcripts for Bmp5 and the Wnt signalling inhibitor *Dkk3* are present at higher levels near the crypt–villus junction than the crypt base (Fig. 3d).

Next, we quantified the density of the transcripts for these signalling mediators in telocytes, subdivided by their position along the crypt–villus axis (Fig. 3e–k). Expression of *Wnt2b* and *Rpso3* was much higher in telocytes at the crypt base than at the villus tip (Fig. 3f, g), while expression of *Dkk3* was increased towards the villus tip (Fig. 3k) and that of *Sfrp1* expression was only marginally dependent on location (Fig. 3j). Expression of *Bmp5* and *Wnt5a* was elevated towards the crypt–villus junction and villus tip, suggesting that the planar cell

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

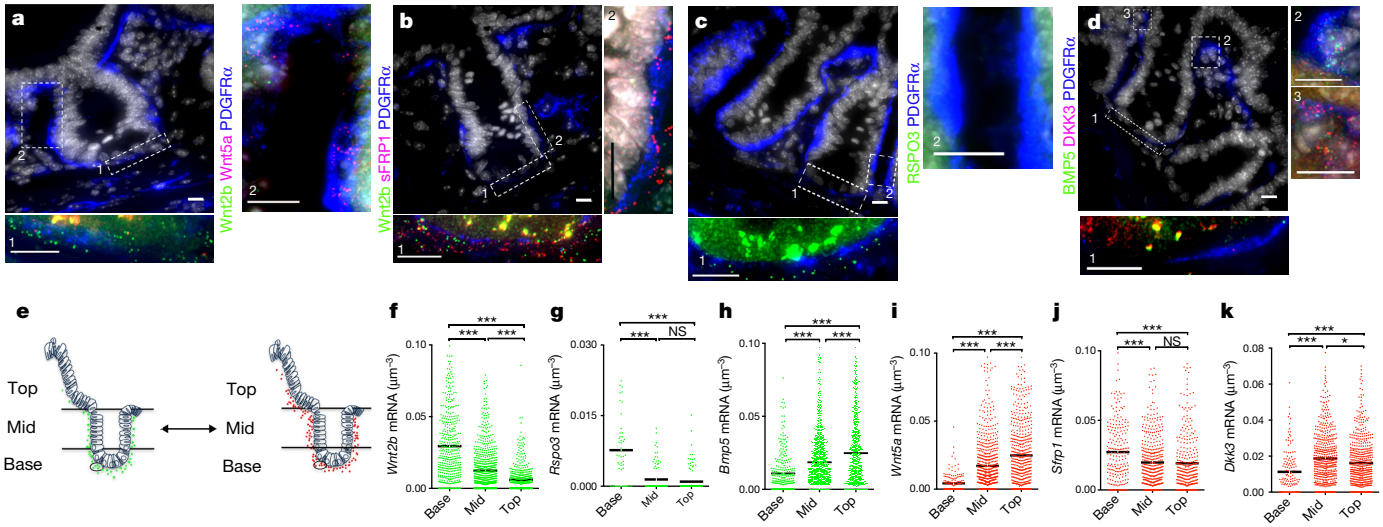


Fig. 3 | Telocytes compartmentalize signalling molecule mRNAs for localized signalling. a–d, Cryosections of mouse duodenum stained for PDGFRα (blue), DNA (grey) and hybridized with smFISH probes as indicated. The numbered insets are shown at higher magnified. **e**, For mRNA quantification, fluorescence signals were counted per defined area along the crypt–villus axis. **f–k**, mRNA molecules per cubic μm along the crypt–villus axis. Experiments were repeated three times with similar results. Scale bars, 10 μm. Black bars represent the mean value. ***$P < 0.001$, *$P < 0.1$; ordinary one-way ANOVA multiple comparison with nonparametric Kruskal–Wallis test. NS, not significant.

polarity pathway and BMP signalling are more active away from the crypt. These data indicate that telocytes compartmentalize transcripts depending upon their position along the crypt–villus axis to locally promote proliferation by maintaining a higher Wnt:SFRP ratio near the crypt base than at the villus epithelium. To obtain independent validation of transcriptome specialization among FOXL1$^+$ telocytes, we also performed single-cell reverse transcription with quantitative PCR (qRT–PCR) for key genes on fluorescence-activated cell sorting (FACS)-sorted *Foxl1-cre;Rosa26-mTmG* GFP-positive cells and identified clearly distinct expression profile subtypes among FOXL1$^+$ cells (Extended Data Fig. 2). Future studies will be needed to determine whether these are temporal states that are dependent on telocyte migration along the radial crypt–villus axis, or whether they represent permanent telocyte subtypes.

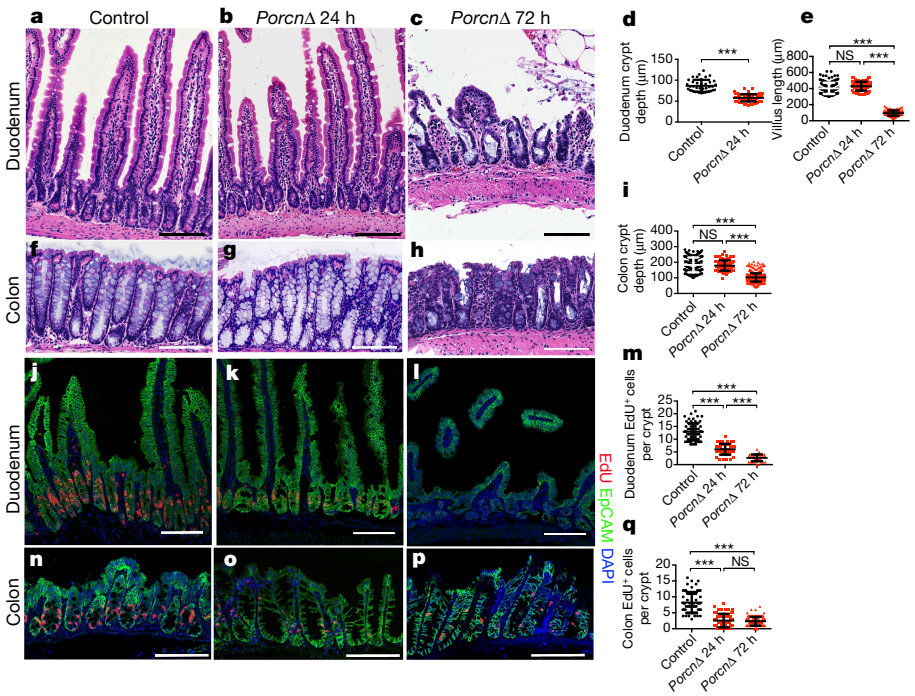


Fig. 4 | The intestinal epithelium depends on Wnts secreted from FOXL1$^+$ telocytes. a–c, Control and *Porcn*Δ mutant duodenum, one and three days after tamoxifen injection. Experiments were repeated at least three times with similar results. **d**, **e**, Duodenal crypt depth (control, $n = 3$ mice; *Porcn*Δ 24 h, $n = 3$ mice) and villus length (control, $n = 9$ mice; *Porcn*Δ 24 h, $n = 3$ mice; *Porcn*Δ 72 h, $n = 7$ mice; ***$P < 0.001$, unpaired two-tailed *t*-test). **f–h**, Control and *Porcn*Δ mutant colon, one and three days after tamoxifen injection. Experiments were repeated at least three times with similar results. **i**, Colonic crypt depth (control, $n = 7$ mice; *Porcn*Δ 72 h, $n = 3$ mice; *Porcn*Δ 72 h, $n = 7$ mice; ***$P < 0.001$, unpaired two-tailed *t*-test). **j–l**, **n–p**, EdU incorporation (red) in the epithelium (EPCAM, green) of duodenum and colon. Experiments were repeated at least three times with similar results. **m**, Number of proliferating cells per duodenal crypt. $n = 3$ mice per group; ***$P < 0.001$, unpaired two-tailed *t*-test. **q**, Number of proliferating cells per colonic crypt. $n = 3$ mice per group; ***$P < 0.001$, unpaired two tailed *t*-test. Centre lines, mean; error bars, s.d. Scale bars, 100 μm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

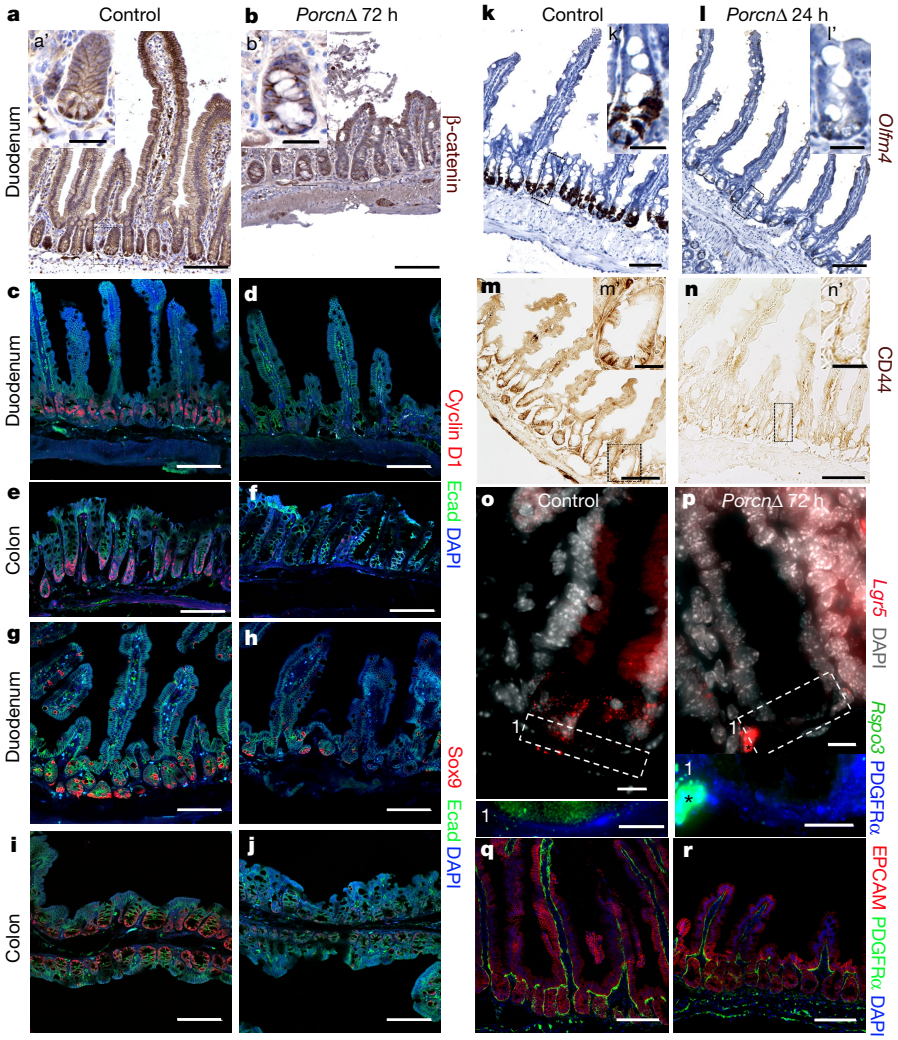


Fig. 5 | FOXL1$^+$ telocytes provide essential Wnt ligands to the intestinal stem cell compartment. a, **b**, Wnt pathway activation analysed by immunohistochemistry for β-catenin (brown). The indicated regions are shown at high magnification in the insets. **c**–**j**, Immunofluorescence of the Wnt targets cyclin D1 and SOX9. **k**–**n**, Stem cell markers *Olfm4* (in situ hybridization) and CD44 (immunohistochemistry), one day post induction in *Porcn*Δ duodenum. Insets show a higher magnification of the indicated regions. **o**, **p**, Expression of *Lgr5* and *Rspo3* mRNA as detected by smFISH, three day post induction in *Porcn*Δ duodenum. *denotes a staining artefact. **q**, **r**, Immunofluorescence for EPCAM and PDGFRα in duodenum of control and *Porcn*Δ mice. Experiments were repeated at least three times with similar results. In **a**–**n**, scale bars, 100 μm (main panel), 25 μm (inset).

To address the question of whether Wnt production by FOXL1$^+$ telocytes is important for intestinal stem cell activity, we derived a *Foxl1-creERT2* mouse line to enable temporally controlled gene ablation in this cell type (Extended Data Fig. 3a). To confirm that the *Foxl1-creERT2* transgene, which is based on the same 170-kb bacterial artificial chromosome (BAC) transgene as we employed previously to obtain faithful reproduction of the *Foxl1* gene expression pattern[10], is active in subepithelial telocytes, we crossed these mice to *Rosa26-mTmG* reporter mice (Extended Data Fig. 3b, c) and treated them with tamoxifen to induce Cre activity. In both small and large bowel, *Foxl1-creERT2* was active in subepithelial telocytes as intended.

Next, we crossed *Foxl1-creERT2* mice with mice carrying a floxed allele of the X-linked *Porcn* gene[15], to obtain *Porcn*$^{loxP/Y}$;*Foxl1-creERT2* mice (hereafter referred to as *Porcn*Δ, Extended Data Fig. 3d). This model enables conditional ablation of all Wnt secretion from FOXL1-expressing telocytes. We analysed *Porcn*Δ mice after 24 and 72 h of tamoxifen treatment. Notably, while epithelial morphology was normal at the 24-h time point (Fig. 4a, b, f, g), stem and progenitor cell proliferation in duodenal and colonic crypts was already markedly reduced, as determined by labelling cells in S-phase with a short pulse of the thymidine analogue ethynyl deoxyuridine (EdU) (Fig. 4j, k, m, n, o, r). Consequently, duodenal crypt depth was significantly reduced even after one day of porcupine ablation (Fig. 4d). After three days of tamoxifen treatment, *Porcn*Δ mice exhibited marked changes in the epithelial architecture in both the small and large intestine (Fig. 4c, h), which we attribute to lack of resupply of the epithelium with newly differentiated cells from the crypts. Epithelial proliferation rates were reduced even further at this time point (Fig. 4l, m, p, r). As a result, the length of the villi was reduced in the duodenum (Fig. 4c, e), and epithelial cells appeared abnormal in colonic crypts (Fig. 4h), which also exhibited a significant reduction in depth (Fig. 4i). Thus, eliminating Wnt secretion from FOXL1$^+$ telocytes results in a large and highly significant reduction in epithelial cell proliferation in the small and large intestine.

In order to test directly whether FOXL1$^+$ telocytes are a relevant source of Wnt signals to the intestinal crypt, we stained control and *Porcn*Δ intestines for the presence of nuclear β-catenin, a marker of active Wnt signalling, as well as cyclin D1 and SOX9, proteins encoded by direct target genes of β-catenin. As shown in Fig. 5a, b, while the control duodenum showed the typical distribution of β-catenin protein at the adherens junctions of all epithelial cells and in the nuclei of cells with active Wnt signalling in the crypts, the *Porcn*Δ intestine had lost

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

nuclear β-catenin protein expression. Likewise, expression of cyclin D1 and SOX9 was reduced markedly in the duodenum and colon of *Porcn*Δ mice (Fig. 5c–j). Indeed, we found that expression of *Olfm4*, *Lgr5* and CD44, all markers of stem cells in the crypt base[16,17], was lost as early as 24 h following porcupine ablation (Fig. 5k–p). Taken together, these findings demonstrate that FOXL$^+$ telocytes are necessary for Wnt pathway activation and proliferation in intestinal stem and progenitor cells. Recently, systemic administration of ligand-specific pharmacological perturbations showed that Wnt and R-spondin ligands represent non-equivalent ligand families that cooperate to maintain intestinal stem cells, with the former having a 'priming' effect and the latter driving stem-cell expansion[18]. Therefore, we evaluated whether Wnt production by FOXL$^+$ telocytes might indirectly influence R-spondin gene expression. Remarkably, we found reduced *Rspo3* mRNA levels in telocytes in the *Porcn*Δ intestine (Fig. 5o, p). Telocytes were still present even after 72 h of *Porcn* gene ablation (Fig. 5q, r), suggesting the possibility that reciprocal signalling from the epithelial compartment might be responsible for maintaining *Rspo3* gene expression. In sum, sub-epithelial FOXL1$^+$ telocytes contribute important Wnt signals to intestinal crypt cells to main their proliferation and allow for homeostatic renewal of the intestinal epithelium.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0084-4.

Received: 11 January 2017; Accepted: 27 February 2018;
Published online 2 May 2018.

1. Sato, T. et al. Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469**, 415–418 (2011).
2. Durand, A. et al. Functional intestinal stem cells after Paneth cell ablation induced by the loss of transcription factor Math1 (Atoh1). *Proc. Natl Acad. Sci. USA* **109**, 8965–8970 (2012).
3. Kim, T. H., Escudero, S. & Shivdasani, R. A. Intact function of Lgr5 receptor-expressing intestinal stem cells in the absence of Paneth cells. *Proc. Natl Acad. Sci. USA* **109**, 3932–3937 (2012).
4. Kabiri, Z. et al. Stroma provides an intestinal stem cell niche in the absence of epithelial Wnts. *Development* **141**, 2206–2215 (2014).
5. San Roman, A. K., Jayewickreme, C. D., Murtaugh, L. C. & Shivdasani, R. A. Wnt secretion from epithelial cells and subepithelial myofibroblasts is not required in the mouse intestinal stem cell niche in vivo. *Stem Cell Reports* **2**, 127–134 (2014).
6. Aoki, R. et al. Foxl1-expressing mesenchymal cells constitute the intestinal stem cell niche. *Cell Mol. Gastroenterol. Hepatol.* **2**, 175–188 (2016).
7. Kaestner, K. H. et al. Six members of the mouse forkhead gene family are developmentally regulated. *Proc. Natl Acad. Sci. USA* **90**, 7628–7631 (1993).
8. Fukuda, K. et al. Mesenchymal expression of Foxl1, a winged helix transcriptional factor, regulates generation and maintenance of gut-associated lymphoid organs. *Dev. Biol.* **255**, 278–289 (2003).
9. Kaestner, K. H. et al. Clustered arrangement of winged helix genes *Fkh-6* and *MFH-1*: possible implications for mesoderm development. *Development* **122**, 1751–1758 (1996).
10. Sackett, S. D., Fulmer, J. T., Friedman, J. R. & Kaestner, K. H. *Foxl1-cre* BAC transgenic mice: a new tool for gene ablation in the gastrointestinal mesenchyme. *Genesis* **45**, 518–522 (2007).
11. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
12. Cantarero Carmona, I., Luesma Bartolomé, M. J. & Junquera Escribano, C. Identification of telocytes in the lamina propria of rat duodenum: transmission electron microscopy. *J. Cell. Mol. Med.* **15**, 26–30 (2011).
13. Cretoiu, D., Cretoiu, S. M., Simionescu, A. A. & Popescu, L. M. Telocytes, a distinct type of cell among the stromal cells present in the lamina propria of jejunum. *Histol. Histopathol.* **27**, 1067–1078 (2012).
14. Vannucchi, M. G., Traini, C., Manetti, M., Ibba-Manneschi, L. & Faussone-Pellegrini, M. S. Telocytes express PDGFRα in the human gastrointestinal tract. *J. Cell. Mol. Med.* **17**, 1099–1108 (2013).
15. Liu, W. et al. Deletion of *Porcn* in mice leads to multiple developmental defects and models human focal dermal hypoplasia (Goltz syndrome). *PLoS One* **7**, e32331 (2012).
16. Gracz, A. D. et al. Brief report: CD24 and CD44 mark human intestinal epithelial cell populations with characteristics of active and facultative stem cells. *Stem Cells* **31**, 2024–2030 (2013).
17. van der Flier, L. G., Haegebarth, A., Stange, D. E., van de Wetering, M. & Clevers, H. OLFM4 is a robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells. *Gastroenterology* **137**, 15–17 (2009).
18. Yan, K. S. et al. Non-equivalence of Wnt and R-spondin ligands during Lgr5$^+$intestinal stem-cell self-renewal. *Nature* **545**, 238–242 (2017).

Acknowledgements We thank A. Rustgi, C. Lengner and Y. Dor for critical reading of the manuscript; M. Sundaram for suggesting the porcupine gene ablation experiment; and all members of the Kaestner laboratory. We acknowledge funding from NIDDK through R37-DK053839. We thank the University of Pennsylvania's Diabetes Research Center (DRC) for the use of the Functional Genomics Core (P30-DK019525) and the Center for Molecular Studies in Digestive and Liver Diseases for the use of the Molecular Pathology and Imaging and Transgenic and Chimeric Mouse Cores (P30-DK050306).

Reviewer information *Nature* thanks L. Samuelson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.S.-C. and K.H.K. came up with the study concept and design, analysed and interpreted the data and wrote and revised the manuscript. B.T., E.E.M. and S.I. acquired and interpreted smFISH data and reviewed the manuscript. Y.J.W., K.J.W. and A.K. acquired and interpreted immunostaining and RNA-seq data.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0084-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0084-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.H.K.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

***Foxl1-creERT2* mice.** *Foxl1-creERT2* mice for tamoxifen-inducible gene ablation of *loxP*-flanked targets were derived by introducing a *creERT2* cassette, cloned from a K18iresCreER-T2 plasmid[19] (Addgene, #44580), into the coding region of the mouse Forkhead box l1 (*Foxl1*) gene in the BAC RP23-446J14 by BAC recombineering as described previously[10]. The targeting primers were: forward, ctccccg gtggagcggagaggctgctgtcgccatgagccacctcttcagtaagccagcagaagcagggtac; reverse, agcactgtggcctcgggaacctgcagaggaaaataagaaaggaagggaatctaggaatctagtgctcccaaaggg tacctgc.

The resulting *Foxl1-creERT2* BAC DNA was linearized and microinjected into fertilized oocytes. The positive transgenic founders were identified by genomic PCR, and the fidelity of transgene expression was tested by crossing *Foxl1-creERT2* transgenic mice to the *Rosa26-mTmG* reporter line[11,20] and visualizing fluorescent protein expression in tissue sections (Extended Data Fig. 3b, c).

***Foxl-CreERT2;PorcnΔ*, *Foxl1-creERT2;Rosa26-mTmG* and *Foxl1-cre; Rosa26-mTmG* mice.** *Foxl1-creERT2* mice were crossed with *Porcn-ex3-7Neo-flox*[15], Rosa26-membrane-targeted dimer tomato protein or membrane-targeted green fluorescent protein (mG)[11], or *Rosa26-YFP*[20] mice to obtain *Foxl1-CreERT2;PorcnΔ*, *Foxl1-creERT2;Rosa26-mTmG*, or *Foxl1-creERT2;Rosa26-YFP* mice. Non-inducible *Foxl1-cre* mice[10] were also crossed to *Rosa26-YFP* and *Rosa26-mTmG* reporter lines (Jackson Laboratories). The animal experiments were performed according to a protocol approved by the University of Pennsylvania Institutional Animal Care and Use Committee.

Tamoxifen treatment. For *Foxl1-creERT2;PorcnΔ* and control (*Foxl1-creERT2;Rosa26-YFP* or *Foxl1-CreERT2;Rosa26mTmG*) mice, tamoxifen (Sigma-Aldrich, Cas#10540-29-1) was dissolved in corn oil at a concentration of 30 mg/ml by shaking overnight at 37 °C and administered intraperitoneally at 75 mg tamoxifen per kg body weight. Mice were injected on three consecutive days and killed on day 4, corresponding to 72 h tamoxifen treatment, or day 2, corresponding to 24 h tamoxifen treatment.

EdU treatment. For EdU treatment, 10 μM EdU solution was injected intraperitoneally two hours before mice were killed, and EdU was detected according to the manufacturer's protocol (Click-iT Plus EdU Alexa Fluor 555 imaging kit C 10638, Molecular Probes).

FACS, RNA isolation and sequencing library preparation. Isolation of FOXL1$^+$ cells using FACS was performed using *Foxl1-cre;Rosa26-YFP* and *Foxl1-cre;Rosa26-mTmG* mice. Small intestines were dissected and washed thoroughly with Hank's balanced salt solution (HBSS) and were incubated in 5 mM EDTA in HBSS for 10 min at 4 °C. Intestinal villi were scraped off using a coverslip and the remaining tissue was cut into small pieces and incubated in 5 mM EDTA and HBSS on ice for 10 min while pipetting to completely remove the remaining epithelium. After vigorous washes, the remaining mesenchymal fraction was incubated with 6 mg/ml Dispase II/0.05% trypsin solution (Sigma-Aldrich, 04942078001) supplemented with 1 U/ml DNaseI (Invitrogen) at 37 °C, until the solution became cloudy and the mesenchyme was dissociated (15 min). Supernatant was filtered through a 70-μm strainer, centrifuged and resuspended with anti CD45–APC and anti EPCAM–APC antibodies (1:300, Biosciences, and 1:100, BioLegend, respectively) in 10%FBS/5 mM EDTA/HBSS for 45 min at room temperature to exclude immune and epithelial cells from sorting. A single-cell suspension was obtained by filtering the supernatant through a 40-μm strainer and washing with 5 mM EDTA in HBSS, before cell sorting using a BD influx instrument (BD Biosciences). For RNA isolation, GFP$^+$ and Tomato$^+$ cells for FOXL1$^+$ and FOXL1$^-$ mesenchymal cells, respectively, were lysed and total RNA was isolated by column purification (Absolutely RNA Nanoprep Kit, Agilent Technologies). cDNA was synthesized and amplified using Ribo-SPIA technology (Ovation RNA-Seq System V2, NuGEN 7102). cDNA was sonicated and an mRNA sequencing library was prepared using the NEBNext RNA library prep kit (New England BioLabs). All libraries were sequenced on the Illumina Hiseq 2500 with 100-bp single-end reads. Median read depth was 27 million reads per sample.

RNA-seq analysis. RNA-seq was aligned to mm9 using RUM[21] with default parameters. For comparison LGR5$^+$ and differentiated enterocyte data sets were used[22]. Differential expression was performed using edgeR[23]. For heatmap analysis, differentially expressed genes were selected by comparisons of FOXL1$^+$ population with the other populations with 1% FDR cut-off and fold change of two. Genes were ranked by their relative fold change.

Single cell sorting and qRT–PCR analysis. Single Foxl1$^+$ cells were sorted into 96-well plates with 5 μl lysis buffer. Cell lysis, reverse transcription, gene-specific preamplification (22 cycles) and downstream quantitative PCR was performed according to the manufacturer's instructions (Fluidigm Biomark HD, 68000088 K1). A total of 179 cells and 39 genes were analysed in the experiment. Outliers were removed using the identifyOutliers function from SINGuLAR software with detection limit set to 24 cycles (Fluidigm PN100-5066 F1). The expression value (Log_2Ex) for each gene was computed according to the SINGuLAR user manual (Fluidigm PN100-5066 F1). For hierarchal clustering, genes were clustered by the Pearson correlation method; cells were measured by their Euclidean distance and clustered by the ward.D method, all implemented by the hclust function in R. To assess the stability of clusters, we applied the clusterboot algorithm from the fpc package to calculate Jaccard coefficient based on 1,000 bootstrapping[24].

Transmission electron microscopy. *Foxl1cre;Rosa26-mTmG* and C57BL/6 control mice were used for immunoelectron microscopy. The ileum was dissected and fixed for 2 h at 4 °C with paraformaldehyde–lysine–periodate (PLP) fixative[25]. After washing with phosphate-buffered saline, the specimens were sectioned with a vibratome to 70 μm thickness and blocked with protein blocking agent (StartingBlock T20 PBS Blocking Buffer, Thermo Scientific #37539) for 30 min at room temperature, and then incubated in the presence of an anti-GFP antibody (1:100, Abcam AB6673) at 4 °C overnight. After washing, the samples were incubated in the presence of horseradish peroxidase (HRP) biotinylated anti-goat secondary antibody, and the signal was developed using 3,3-diaminobenzidine tetrahydrochloride (DAB) as a substrate. The ultra-thin sections were examined under a JEOL 1010 electron microscope fitted with a Hamamatsu digital camera and AMT Advantage image capture software.

Histology. For haematoxylin and eosin staining, paraffin tissue sections were rehydrated, incubated in haematoxylin for 5 min, rinsed in water, dipped quickly in 1% acid ethanol and washed in water. Tissues then were immersed in 0.2% ammonium hydroxide, rinsed in water, dipped in eosin for 40 s and briefly rinsed in water before dehydration and mounting.

Imunofluorescence and immunohistochemistry. Mouse intestines were rinsed in PBS and fixed with 4% paraformaldehyde overnight, rinsed in PBS, and either dehydrated for paraffin embedding or immersed in 30% sucrose for 4 h at 4 °C, embedded in optimum cutting temperature compound (OCT, Tissue-Tek), and frozen for cryosectioning. Antigen retrieval was performed using citrate buffer pH 6.0 with a pressure cooker (PickCell Laboratories, Agoura Hills, CA). FOXL1 staining required the use of cryosections, antigen retrieval, and amplification of signal using tyramide (TSA systems, PerkinElmer) as previously described[6]. CD44 immunohistochemistry was performed as described[26].Antibodies used were as follows: rabbit SOX9 (1:300; Millipore AB5535), guinea pig FOXL1 (1:1,500), goat GFP (1:200; Abcam AB6673), mouse E-cadherin (1:250; BD Transduction 610182), mouse β-catenin (1:200; BD Transduction 610153), rabbit EPCAM (1:100; Abcam AB71916), goat PDGFRα (1:100; R&D AF1062), rat CD44var (v6) (1:100; Biosciences BMS145). Cy2-, Cy3- and Cy5-conjugated donkey secondary antibodies were purchased from Jackson ImmunoResearch Laboratories. For immunohistochemistry, horseradish peroxidase-conjugated antibodies were incubated for 2 h at room temperature and 3,3-diaminobenzidine tetrahydrochloride (DAB) was used as a substrate to develop the signal. For the studies requiring quantification, we used $n = 3$ per genotype or condition based on the severity of the phenotype observed in pilot experiments. Mice were assigned to condition by genotype and were age-matched from the same colony. The investigator counting the number of EdU-positive cells or quantifying villus length or crypt depth was blinded to the genotype of the mice.

Clearing of adult mouse intestine. *Foxl1-cre; Rosa26-mTmG* and C57BL/6 adult mice were anaesthetized, intestine was dissected and fixed in 4%PFA for 24 h at 4 °C. Following washes with PBS for another 24 h at 4 °C, the intestine was incubated in hydrogel solution for 24 h at 4 °C[27]. After 3 h at 37 °C, the hydrogel-embedded intestine was placed in an X-CLARITY ETC chamber (LOGOS Biosystems) for electrophoretic tissue clearing for 7 h. The cleared intestine was immunostained with goat anti-GFP (1:100; Abcam) or PDGFRα (1:100; R&D) and APC-conjugated anti-mouse EPCAM (1:100; Biolegend) or rabbit anti-EPCAM (1:100; Abcam). For FOXL1 antibody staining, following clearing, antigen retrieval was performed (as described above) in citrate buffer. Primary and secondary antibodies were incubated for 48 h at 4 °C each. The stained intestine was placed en bloc on an image slide using 1-mm depth adhesive silicone isolator, mounted in X-CLARITY mounting solution and imaged using confocal scanning. *Z*-stacks projections were compiled using Volocity software.

Single-molecule FISH and immunofluorescence. Libraries of 48 probes each were designed against the mouse *Wnt2b*, *Wnt5a*, *Sfrp1*, *Rspo3*, *Bmp5* and *Dkk3* mRNAs (Stellaris RNA FISH probes, Biosearch Technologies). Hybridization conditions and imaging were as described previously[28,29], except for the addition of immunofluorescence detection of PDGFRα. For immunofluorescence, a goat anti-PDGFRα antibody was diluted in hybridization buffer (1:100), added to the FISH probes and incubated overnight at 30 °C. Secondary antibody Cy2-conjugated donkey anti-goat (1:200) was added to glucose oxidase (GLOX) buffer for 20 min in room temperature. smFISH imaging was performed on a Nikon-Ti-E inverted fluorescence microscope with a 100 × oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software as previously described[28,30]. Single-cell segmentation and automatic transcript quantification were performed in a custom Matlab program (MATLAB release 2013b, Mathworks) using

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

TransQuant software[31]. Images were filtered with a Laplacian of Gaussian filter of size 15 pixels and standard deviation of 1.5 pixels[32].

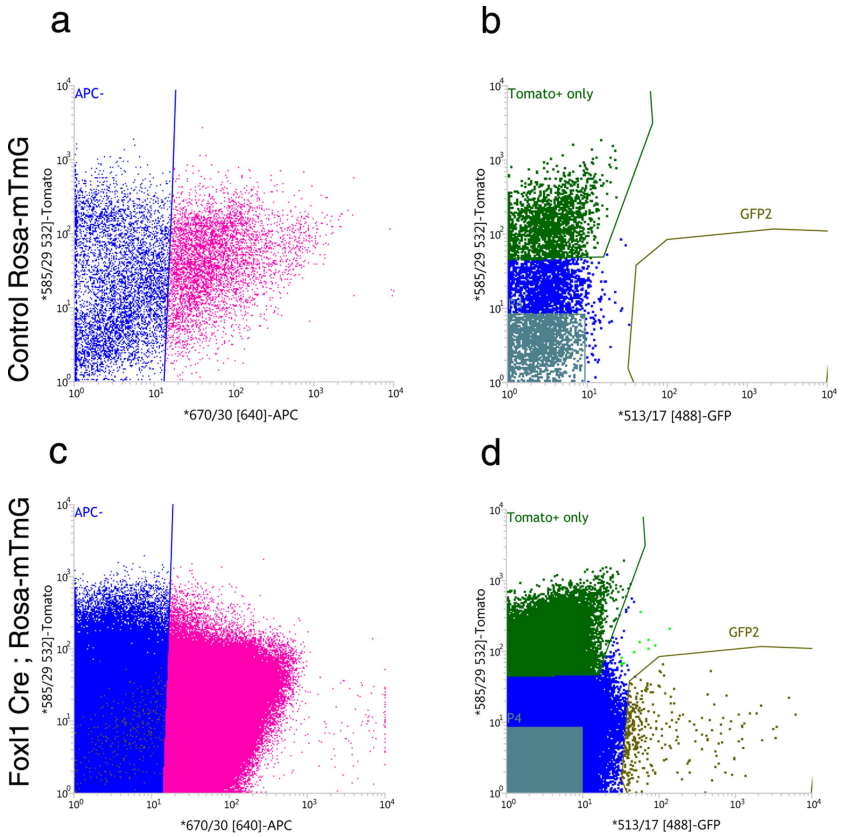
***Olfm4* mRNA in-situ hybridization.** In situ hybridization for Olfm4 mRNA was performed using the RNAscope kit (Advanced Cell Diagnostics) according to the manufacturer's instructions.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. RNA-seq data that support the findings of this study have been deposited in the NCBI GEO under accession code GSE94072. Source data for Figs. 2, 3 and 4 are provided in the online version of the paper.

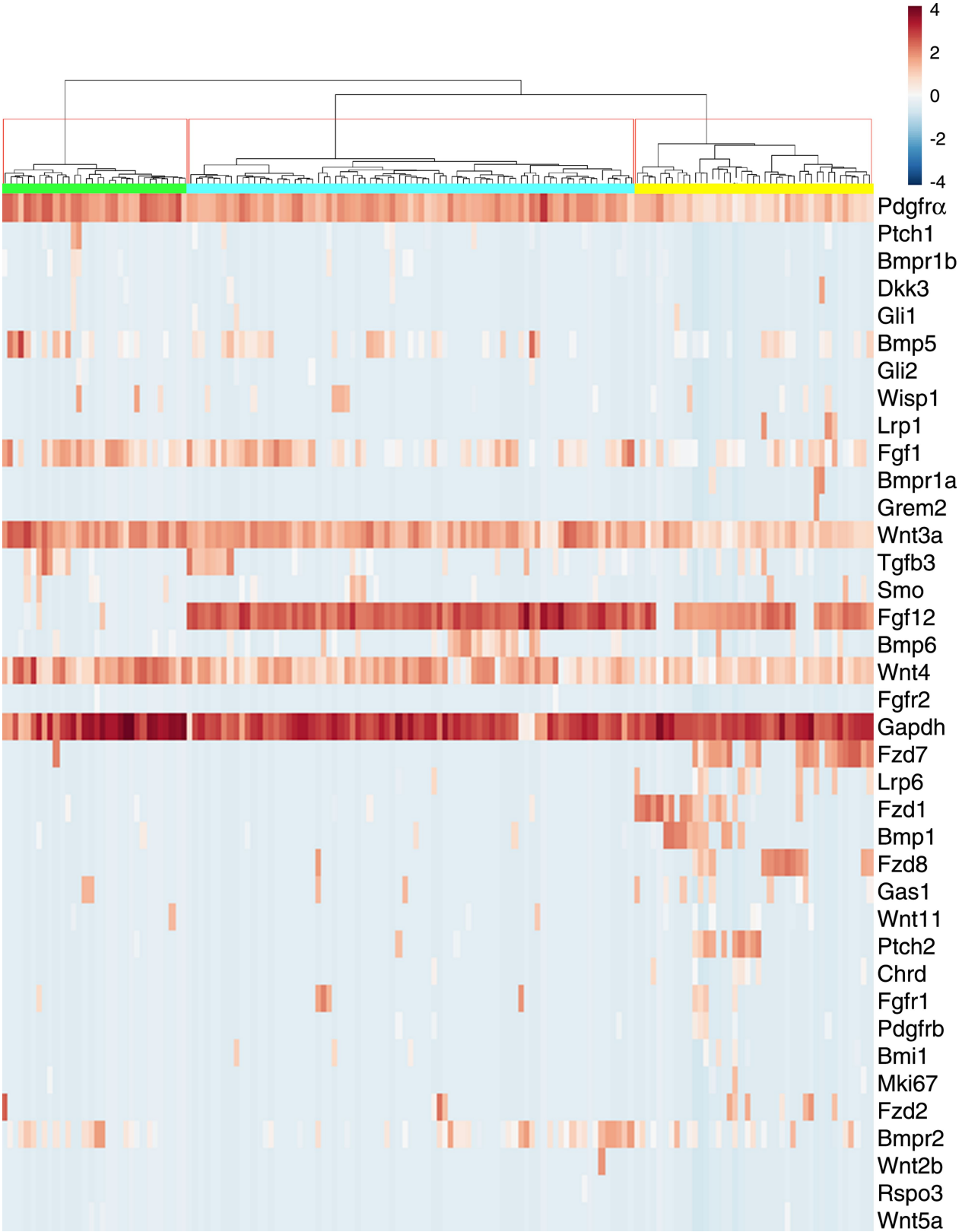
19. Wen, F. et al. Expression of conditional Cre recombinase in epithelial tissues of transgenic mice. *Genesis* **35**, 100–106 (2003).
20. Srinivas, S. et al. Cre reporter strains produced by targeted insertion of EYFP and ECFP into the ROSA26 locus. *BMC Dev. Biol.* **1**, 4 (2001).
21. Grant, G. R. et al. Comparative analysis of RNA-seq alignment algorithms and the RNA-seq unified mapper (RUM). *Bioinformatics* **27**, 2518–2528 (2011).
22. Sheaffer, K. L. et al. DNA methylation is required for the control of stem cell differentiation in the small intestine. *Genes Dev.* **28**, 652–664 (2014).
23. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
24. Hennig, C. Cluster-wise assessment of cluster stability. *Comput. Stat. Data Anal.* **52**, 258–271 (2007).
25. McLean, I. W. & Nakane, P. K. Periodate-lysine-paraformaldehyde fixative. A new fixation for immunoelectron microscopy. *J. Histochem. Cytochem.* **22**, 1077–1083 (1974).
26. Shyer, A. E., Huycke, T. R., Lee, C., Mahadevan, L. & Tabin, C. J. Bending gradients: how the intestinal stem cell gets its home. *Cell* **161**, 569–580 (2015).
27. Chung, K. et al. Structural and molecular interrogation of intact biological systems. *Nature* **497**, 332–337 (2013).
28. Itzkovitz, S. & van Oudenaarden, A. Validating transcripts with probes and imaging technology. *Nat. Methods* **8** (Suppl.), S12–S19 (2011).
29. Lyubimova, A. et al. Single-molecule mRNA detection and counting in mammalian tissue. *Nat. Protoc.* **8**, 1743–1758 (2013).
30. Bahar Halpern, K. et al. Bursty gene expression in the intact mammalian liver. *Mol. Cell* **58**, 147–156 (2015).
31. Bahar Halpern, K. & Itzkovitz, S. Single molecule approaches for quantifying transcription and degradation rates in intact mammalian tissues. *Methods* **98**, 134–142 (2016).
32. Itzkovitz, S. et al. Single-molecule transcript counting of stem-cell markers in the mouse intestine. *Nat. Cell Biol.* **14**, 106–114 (2011).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



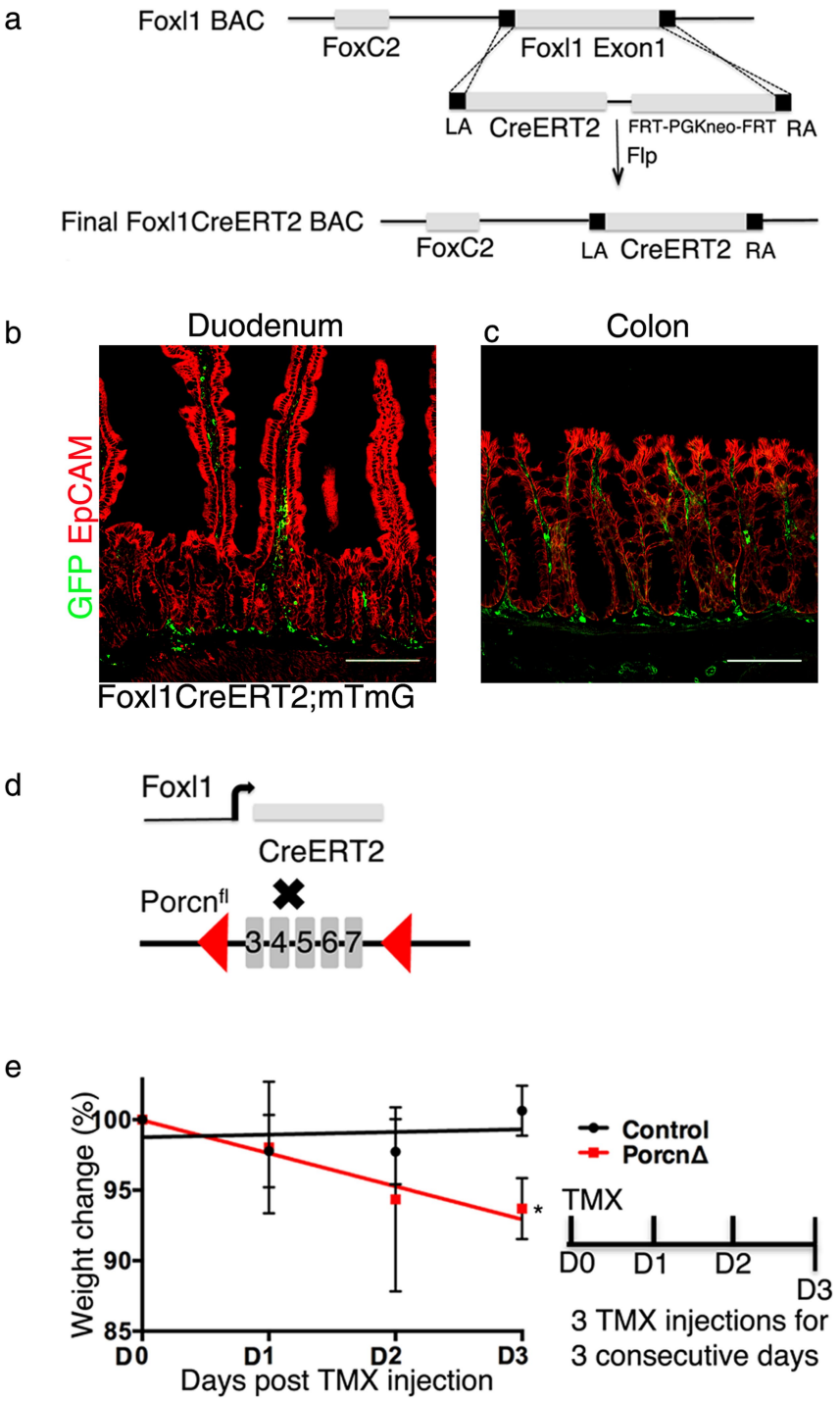
Extended Data Fig. 1 | Fluorescence-activated cell sorting plots. **a**–**d**, Representative FACS plots of mesenchymal cells isolated from *Foxl1-cre;Rosa26-mTmG* (**c**, **d**) as compared to control *Rosa26-mTmG* (**a**, **b**) mice, showing sorting strategy. In **a** and **c**, Allophycocyanin^{+} (APC^{+}), CD45^{+} and EPCAM^{+}-labelled cells were gated out to exclude immune and epithelial cell contamination. In **d**, FOXL1^{+}, GFP^{+} and FOXL1^{-}Tomato^{+} cells were gated and sorted based on fluorescence activity as compared to the negative control (**b**). Experiments were repeated at least three times with similar results.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 2 | Single-cell qPCR of FOXL1$^+$ telocytes showing heterogeneity within the FOXL1$^+$ cell population. Hierarchical clustering of FOXL1$^+$ single cells based on qPCR-based detection of 30 genes. Note that all FOXL1$^+$ cells express *Pdgfra* and *Wnt3a*. The population is clustered into three main groups. Jaccard coefficients for the three clusters were: 0.83 (green), 0.69 (turquoise) and 0.79 (yellow), respectively, indicating underlying cluster stability. Values between 0.6 and 0.75 indicate that the cluster is measuring a pattern in the data. Clusters with stability values above about 0.85 are considered to be highly stable.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 3 | Derivation of *Foxl1-creERT2;PorcnΔ* mice. **a**, Schema for the generation of *Foxl1-creERT2* mice using BAC recombineering. The coding sequence of exon 1 of *Foxl1* was targeted by the sequence of *creERT2*. FRT, flipase recognition target; Flp, flipase; LA, left homology arm; RA, right homology arm. **b**, **c**, Tamoxifen induction of *Foxl1-creERT2;Rosa26-mTmG*-driven expression of membrane-bound GFP to mesenchymal telocytes in the duodenum (**b**) and colon (**c**). Immunofluorescence staining for GFP (green), EPCAM (red). **d**, Schema for the generation of *Foxl1-creERT2;PorcnΔ* mice. *Foxl1-creERT2* mice were crossed with mice carrying *loxP* sites flanking exons 3–7 of the X-linked porcupine homologue (*Porcn*) gene. **e**, *Foxl1-creERT2;Rosa26-mTmG* (control, $n = 5$ mice) and *Foxl1-creERT2;PorcnΔ* (*PorcnΔ*, $n = 8$ mice) male mice were treated with tamoxifen for three consecutive days to induce Cre expression and weighed every day. The slope of the weight loss was significantly different in *PorcnΔ* mice as compared to control mice (*$P = 0.0107$, two-tailed linear regression analysis).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Wnt pathway gene expression levels in different cell populations as assessed by RNA-seq (mean of 2–3 samples each in FPKM)

Wnt pathway

Gene name	Foxl1$^+$ Expression level mean (FPKM)	Mes Expression level mean (FPKM)	Lgr5 Expression level mean (FPKM)	Diff Expression level mean (FPKM)
Wnt2	0.9742	0	0.1736	0
Wnt2b	9.5831	0.1205	0.1559	0.1191
Wnt3	0.3438	0	1.0915	0.0909
Wnt4	1.7951	0	0.0102	0.9198
Wnt5a	4.9932	0	0.1219	0.0786
Wnt5b	0.8935	0	0.5514	0.3542
Wnt6	0.0866	0	0.0824	0.3045
Wnt9a	0.5072	0	0.0505	0.0553
Wisp1	2.4082	0	0.0286	0
Rspo2	0.4112	0	0	0
Rspo3	2.1902	13.3906	0.0141	0
Lgr4	7.5384	0	48.6203	9.8020
Lgr5	1.9220	0	19.3406	0.0408
Lgr6	0.2264	0	0	0
Znrf1	4.2325	0	3.6232	0.8523
Znrf2	2.2594	5.0671	12.7731	7.7214
Znrf3	0.9130	0.1328	1.9908	0.1333
Dkk2	1.8779	0	0	0
Dkk3	36.5673	22.1995	0.0017	0.0077
Wif1	1.2579	0	0.0125	0.0844
Sfrp1	47.2422	0.6198	0.0722	0.0034
Sfrp4	0.7134	0	0.0096	0
Fzd2	2.50329	0	1.9856	0.0100
Fzd4	4.5451	0	0.2025	0.9071
Fzd6	0.5785	0	2.7795	0.2181
Fzd7	1.1146	0	16.7951	2.5999
Fzd8	0.6245	0	0.7510	0.3280
Lrp1	11.4018	0.0144	6.4567	12.5752
Lrp6	6.0929	0.0013	9.5234	4.2674
Tcf4	15.0228	8.338	6.0325	0.9127
Tcf15	1.71113	0	0.0090	0
Tcf19	1.456	8.768	13.6231	0.7681
Tcf21	19.6602	0.0065	0	0
Tcf7l1	1.8914	0	0.1070	0.0623

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0075-5

De novo formation of the biliary system by TGFβ–mediated hepatocyte transdifferentiation

Johanna R. Schaub[1,12], Kari A. Huppert[2,12], Simone N. T. Kurial[1,3,12], Bernadette Y. Hsu[1,3], Ashley E. Cast[2], Bryan Donnelly[4], Rebekah A. Karns[2], Feng Chen[1], Milad Rezvani[1], Hubert Y. Luu[5], Aras N. Mattis[6,7], Anne-Laure Rougemont[8], Philip Rosenthal[7,9], Stacey S. Huppert[2,10,13]* & Holger Willenbring[1,7,11,13]*

Transdifferentiation is a complete and stable change in cell identity that serves as an alternative to stem-cell-mediated organ regeneration. In adult mammals, findings of transdifferentiation have been limited to the replenishment of cells lost from preexisting structures, in the presence of a fully developed scaffold and niche[1]. Here we show that transdifferentiation of hepatocytes in the mouse liver can build a structure that failed to form in development—the biliary system in a mouse model that mimics the hepatic phenotype of human Alagille syndrome (ALGS)[2]. In these mice, hepatocytes convert into mature cholangiocytes and form bile ducts that are effective in draining bile and persist after the cholestatic liver injury is reversed, consistent with transdifferentiation. These findings redefine hepatocyte plasticity, which appeared to be limited to metaplasia, that is, incomplete and transient biliary differentiation as an adaptation to cell injury, based on previous studies in mice with a fully developed biliary system[3–6]. In contrast to bile duct development[7–9], we show that de novo bile duct formation by hepatocyte transdifferentiation is independent of NOTCH signalling. We identify TGFβ signalling as the driver of this compensatory mechanism and show that it is active in some patients with ALGS. Furthermore, we show that TGFβ signalling can be targeted to enhance the formation of the biliary system from hepatocytes, and that the transdifferentiation-inducing signals and remodelling capacity of the bile-duct-deficient liver can be harnessed with transplanted hepatocytes. Our results define the regenerative potential of mammalian transdifferentiation and reveal opportunities for the treatment of ALGS and other cholestatic liver diseases.

In regenerating organs of adult mammals, differentiated cells can replenish other types of differentiated cells by transdifferentiation, as in the pancreatic islet[10], gastric gland[11], lung alveolus[12] and intestinal crypt[13]. Whether mammalian transdifferentiation can build these or other structures de novo is unknown. In the liver, hepatocytes can undergo biliary differentiation to form reactive ductules both in humans[14–17] and animals with cholestatic liver injury[3–6,18–21]. However, hepatocyte-derived biliary cells exhibit incomplete biliary and residual hepatocyte differentiation, that is, they are not mature cholangiocytes, and revert back to their original identity after the injury is reversed[3–6]. Moreover, hepatocyte-derived ductules do not contribute to bile drainage[5]. These findings are consistent with metaplasia, but not with transdifferentiation, and call into question the functional importance of hepatocyte plasticity. However, all studies of hepatocyte plasticity published so far used animals with a fully developed biliary system in which residual cholangiocytes are available to regenerate injured bile ducts, probably leading to insufficient pressure for transdifferentiation. To determine the full extent of hepatocyte plasticity, we used mice that lack the intrahepatic biliary system.

Specifically, we used mice that mimic the hepatic phenotype of ALGS—a human disease caused by impaired NOTCH signalling[22–24]—generated by deletion of floxed alleles of the NOTCH signalling effector RBPJ and, to prevent compensation[25], the transcription factor HNF6 (also known as ONECUT1) in embryonic liver progenitors using Cre expressed from an albumin (*Alb*) promoter[2]. These *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice are severely cholestatic because they lack peripheral bile ducts (pBDs), the branches of the intrahepatic biliary tree, at birth; however, more than 90% of the mice survive and form pBDs and a functional biliary tree by postnatal day (P) 120[26] (Fig. 1a). Although *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice have hilar bile ducts (hBDs), the trunk of the biliary tree, we hypothesized that the new pBDs originate from hepatocytes because the livers of *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice contain hybrid cells that express hepatocyte and biliary markers[2,26]. To test this hypothesis, we developed Cre-independent hepatocyte fate tracing. For this, we activated the Flp-reporter (GFP) in *R26ZG*$^{+/+}$ mice specifically in hepatocytes by intravenous injection of an adeno-associated virus (AAV) serotype 8 vector expressing Flp from the transthyretin (*Ttr*) promoter (AAV8-Ttr-Flp) (Fig. 1a and Extended Data Fig. 1a–g).

To trace hepatocytes during pBD formation, we intravenously injected *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice with AAV8-Ttr-Flp at P17, which is before cells expressing biliary markers emerge in the liver periphery[2,26]. At P120, the newly formed pBDs, identified by wide-spectrum (ws) cytokeratin (CK) immunofluorescence, were GFP positive, indicating hepatocyte origin (Fig. 1a, b). By contrast, pBDs that formed normally during liver development in Cre-negative littermates were GFP negative (Extended Data Fig. 2a, b). In many pBDs of *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice, all cells were GFP positive, whereas the overall labelling efficiency of peripheral cholangiocytes was 39.2 ± 7.2% (mean ± s.e.m.); however, this number correlated well with the hepatocyte labelling efficiency (36.6 ± 4.8%; $P = 0.48$, paired two-sided Student's *t*-test) (Extended Data Fig. 2c), indicating that all peripheral cholangiocytes originated from hepatocytes. We also observed hepatocyte-derived peripheral cholangiocytes in mice in which hepatocytes were labelled after weaning (P39) (Extended Data Fig. 2d, e). These results show that hepatocytes can form pBDs de novo.

To investigate the function of the hepatocyte-derived pBDs (HpBDs), we determined whether they are contiguous with the extrahepatic biliary system. At P30, retrograde ink injection into the common bile duct filled only hBDs, reflecting a severely truncated biliary tree that lacked pBDs; however, ink injection at ≥P120 revealed a biliary tree of

[1]Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California San Francisco, San Francisco, CA, USA. [2]Division of Gastroenterology, Hepatology and Nutrition, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [3]Biomedical Sciences Graduate Program, University of California San Francisco, San Francisco, CA, USA. [4]Department of Pediatric General and Thoracic Surgery, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [5]Department of Surgery, Division of General Surgery, University of California San Francisco, San Francisco, CA, USA. [6]Department of Pathology, University of California San Francisco, San Francisco, CA, USA. [7]Liver Center, University of California San Francisco, San Francisco, CA, USA. [8]Department of Genetic and Laboratory Medicine, Division of Clinical Pathology, Geneva University Hospital, Geneva, Switzerland. [9]Department of Pediatrics, Division of Gastroenterology, Hepatology and Nutrition, University of California San Francisco, San Francisco, CA, USA. [10]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. [11]Department of Surgery, Division of Transplant Surgery, University of California San Francisco, San Francisco, CA, USA. [12]These authors contributed equally: Johanna R. Schaub, Kari A. Huppert, Simone N. T. Kurial. [13]These authors jointly supervised this work: Stacey S. Huppert, Holger Willenbring *e-mail: holger.willenbring@ucsf.edu; stacey.huppert@cchmc.org

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

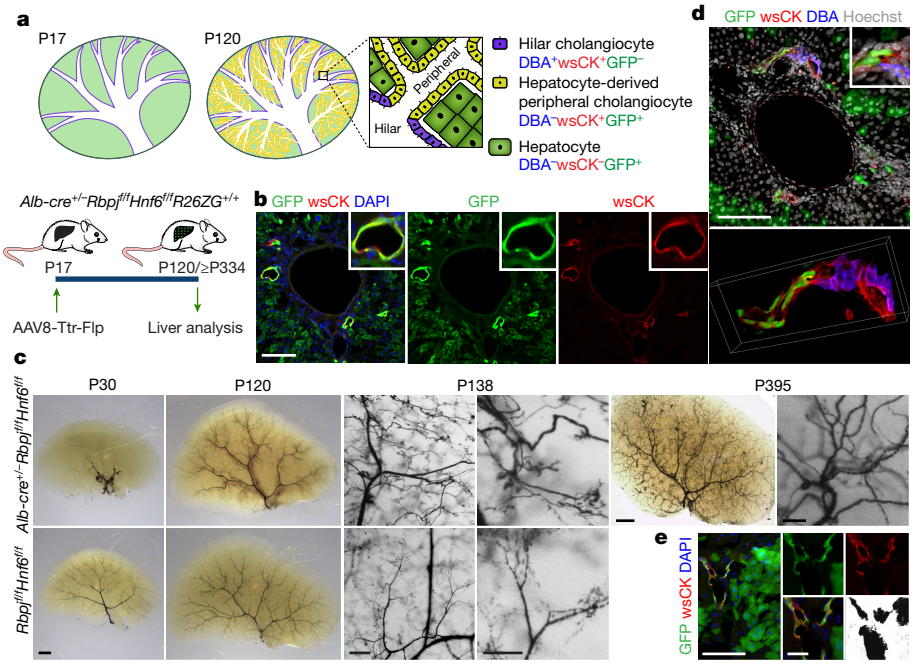


Fig. 1 | Hepatocytes can convert into peripheral cholangiocytes and form pBDs contiguous with preexisting hBDs. a, De novo pBD formation and hepatocyte fate tracing in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice. Cells identified by dolichos biflorus agglutinin (DBA) lectin labelling and wsCK and GFP immunofluorescence. **b**, Immunofluorescence of hepatocyte-fate-traced P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mouse liver ($n = 7$). **c**, Biliary tree visualized by retrograde ink injection into the common bile duct of P30 ($n = 6$), P120–P138 ($n = 6$) and ≥P334 ($n = 6$) *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice, and P30 ($n = 3$) and P120–P138 ($n = 5$) *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice. **d**, Maximum projection (top) and 3D reconstruction (bottom) of *z*-stack image of hepatocyte-fate-traced P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mouse liver ($n = 2$). **e**, Immunofluorescence and bright-field images of hepatocyte-fate-traced P468 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mouse liver after retrograde ink injection into the common bile duct (≥P334, $n = 3$). Scale bars, 2 mm (**c**, P30, P395 left), 500 μm (**c**, P138 left), 250 μm (**c**, P138 right), 100 μm (**b**, **c**, P395 right, **d**, **e**, left), 25 μm (**e**, middle).

normal dimensions, demonstrating that HpBDs are connected to hBDs (Fig. 1c). Three-dimensional (3D) reconstruction of confocal imaging confirmed that HpBDs form contiguous lumens with dolichos biflorus agglutinin (DBA)-labelled hBDs (Fig. 1d and Supplementary Video 1). Accordingly, HpBDs were effective in draining bile, as evidenced by normalization of serum markers of cholestasis (total bilirubin and alkaline phosphatase levels; Extended Data Fig. 3a–c) and hepatocyte injury (alanine aminotransferase and aspartate aminotransferase levels; Extended Data Fig. 3d, e) and resolution of cholestasis-induced liver fibrosis (Sirius red staining; Extended Data Fig. 3f). A few mice that continued to have increased levels of serum total bilirubin and liver fibrosis showed abundant wsCK-positive hepatocytes but no HpBDs (Extended Data Fig. 3a, b, g). In contrast to the transient hepatocyte-derived ductules observed in other mouse models of cholestatic liver injury[3–6], HpBDs were stable beyond the time when cholestasis and liver injury were resolved and were maintained for life (≥P334) (Fig. 1c, e). These results show that HpBDs provide normal and stable biliary function.

We also investigated the authenticity and maturity of the cholangiocytes that form the HpBDs. The cells displayed acetylated tubulin-marked primary cilia (Fig. 2a), indicating a switch from hepatocyte to biliary fate[19], and somatostatin receptor 2 (SSTR2) (Fig. 2b), a marker of biliary function[27]. They also expressed the mature biliary markers epithelial cell adhesion molecule (EPCAM) and CK19 (Fig. 2c), which are lacking or expressed at low levels in hepatocyte-derived metaplastic biliary cells[3,5]. To substantiate these results, we performed RNA-sequencing (RNA-seq) analysis on hepatocyte-derived peripheral cholangiocytes isolated as EPCAM-positive DBA-negative cells by FACS from >P115 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice (Extended Data Fig. 4a). *Rbpj* and *Hnf6* were deleted in these cells (Extended Data Fig. 4b, c), which was expected because *Alb-cre* is active in embryonic liver progenitors before they commit to hepatocyte or biliary fate[2,9]. Principal component analysis showed that hepatocyte-derived peripheral cholangiocytes cluster closely with normal peripheral cholangiocytes isolated from Cre-negative littermates, but not with parental hepatocytes (Fig. 2d, e and Supplementary Table 1). Accordingly, hepatocyte-derived peripheral cholangiocytes expressed normal levels of mature biliary markers that are virtually undetectable in 3,5-diethoxycarbonyl-1,4-dihydrocollidine (DDC) diet-induced hepatocyte-derived metaplastic biliary cells[3], except for CFTR, which is less enriched in cholangiocytes than the other markers (Fig. 2f). Hepatocyte-derived peripheral cholangiocytes also expressed other commonly used markers of biliary differentiation[3,5,20,25,27,28] at normal or near-normal levels (Fig. 2f), and showed virtually no memory of hepatocyte differentiation[3,29,30] (Fig. 2g and Extended Data Fig. 4d). These results show that hepatocyte-derived peripheral cholangiocytes are authentic and mature peripheral cholangiocytes.

Next, we investigated the contribution of proliferation to HpBD formation, and sought to determine whether a few hepatocytes proliferate extensively after transdifferentiation, or whether many hepatocytes transdifferentiate. For this, we measured clonal expansion in HpBDs by sparsely labelling hepatocytes in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice with low-dose AAV8-Ttr-Flp at P18 and quantifying the cells in GFP-positive clones at P120 (Fig. 3a). Surprisingly, we found only 1.56 ± 0.05 cells per clone in HpBDs, with one-cell clones accounting for 63.8 ± 3.6% of clones. To exclude the possibility that the nonintegrating AAV vector missed proliferating hepatocytes, we analysed clone size in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26R-Confetti*$^{+/-}$ mice in which hepatocytes and hilar cholangiocytes are stochastically labelled with one of four fluorescent proteins (Fig. 3b and Supplementary Video 2). We found 1.91 ± 0.06 cells per clone and 45.6 ± 1.1% one-cell clones in HpBDs at P150. In addition, the number of cells per clone in hBDs was similar between these mice and *Alb-cre*$^{+/-}$*R26R-Confetti*$^{+/-}$ control mice at P90 (2.08 ± 0.08 and 2.41 ± 0.01; $P = 0.060$, two-sided Student's *t*-test) (Extended Data Fig. 5a). These results show that HpBDs form with little proliferation

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

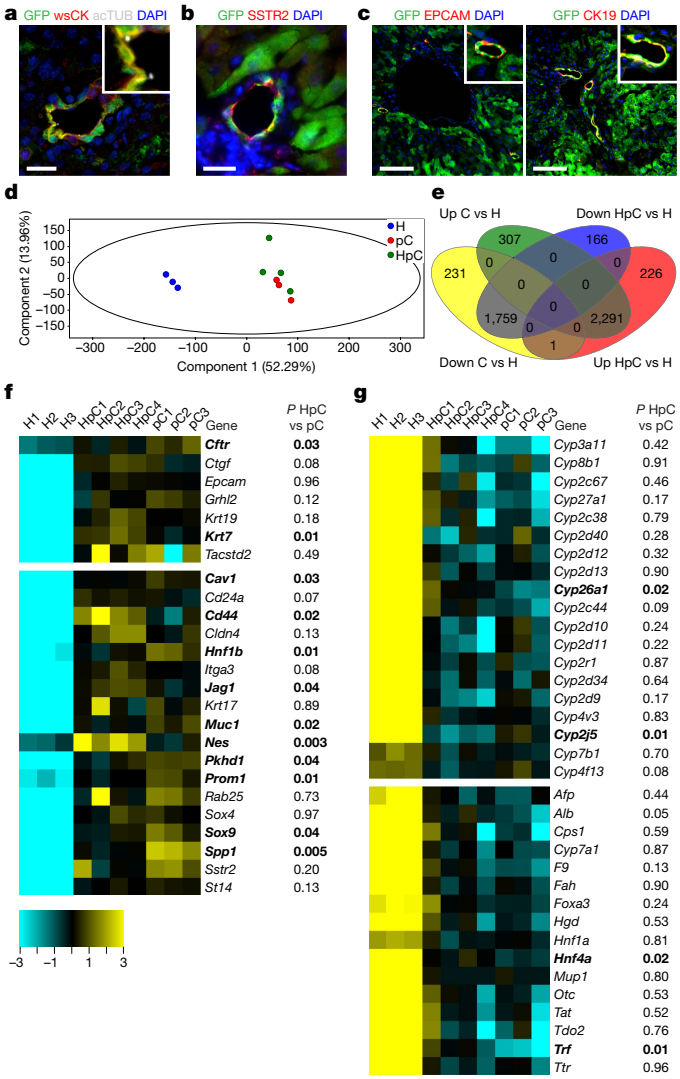


Fig. 2 | Hepatocyte-derived peripheral cholangiocytes are equivalent to normal mature peripheral cholangiocytes. a–c, Immunofluorescence of hepatocyte-fate-traced P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mouse liver ($n = 3$ each). acTUB, acetylated tubulin. Scale bars, 100 μm (**c**), 20 μm (**a**, **b**). **d–g**, RNA-seq analysis of normal peripheral cholangiocytes (pC; $n = 3$ mice), hepatocyte-derived peripheral cholangiocytes (HpC; $n = 4$ mice) and RBPJ- and HNF6-deficient hepatocytes (H; $n = 3$ mice). **d**, Principal component analysis. **e**, Venn diagram showing the number of genes significantly differentially up- and downregulated in pC or HpC versus H. **f**, Heat maps of genes reflecting cholangiocyte differentiation, including genes lacking in DDC diet-induced hepatocyte-derived metaplastic biliary cells (top) and other marker genes (bottom). **g**, Heat maps of genes reflecting hepatocyte differentiation, including all differentially expressed cytochrome P450 (CYP) genes enriched in adult mouse liver (top) and other marker genes (bottom). One-way ANOVA, false-discovery rate (FDR)-corrected $P < 0.05$; fold change > 3 (**e–g**). Two-sided Student's *t*-test; bold genes $P < 0.05$ for HpC versus pC (**f**, **g**).

and are thus polyclonal, and confirm that hilar cholangiocytes do not contribute to HpBD formation.

Unlike in HpBDs, we found significant proliferation in hepatocyte-derived reactive ductules, which are detectable in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice during cholestasis (Extended Data Fig. 5b). Clonal analysis showed 2.70 ± 0.08 cells per clone in reactive ductules in P90 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26R-Confetti*$^{+/-}$ mice, with more three- and four-cell clones than in pBDs in *Alb-cre*$^{+/-}$*R26R-Confetti*$^{+/-}$ mice (Extended Data Fig. 5c). Because HpBDs form with little proliferation, we reasoned that these proliferating cells are hepatocyte-derived metaplastic biliary cells[3–6]. Indeed, we found that 93.51 ± 0.81% of the proliferating biliary cells identified by KI67 and osteopontin (OPN) immunofluorescence in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice near the peak of cholestasis (P54) were CK19 negative (Extended Data Fig. 5d).

We also investigated proliferation in established HpBDs by inducing cholestatic liver injury with a DDC diet in >P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice. Reactive ductules formed 2 weeks later than in Cre-negative littermates, and consisted mainly of OPN-positive EPCAM-negative cells, indicative of hepatocyte metaplasia (Extended Data Fig. 5e). Indeed, after feeding a DDC diet to *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice in which hepatocyte fate tracing was induced at >P120, we found that 84.5 ± 5.6% of the cells in reactive ductules originated from hepatocytes, in contrast to 20.1 ± 3.0% in Cre-negative littermates (Extended Data Fig. 5f). These results confirm that NOTCH signalling is important for cholangiocyte proliferation[31], which explains our finding of a limited role of proliferation in HpBD formation and underscores the authenticity of hepatocyte-derived peripheral cholangiocytes.

As illustrated by a lack of pBDs in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice (Fig. 1c), NOTCH signalling is needed for bile duct development[7–9], raising the question of which signalling pathway drives HpBD formation in its absence (Extended Data Fig. 4b, c). We focused on TGFβ signalling because it promotes biliary differentiation and morphogenesis in development[32], although it is not essential, as shown by the development of a normal biliary tree in mice that lack the TGFβ type II receptor (TGFBR2) in embryonic liver progenitors (Extended Data Fig. 6a). We reasoned that TGFβ signalling is induced in hepatocytes of *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice because of the paradoxical role of HNF6 in liver development—it activates the biliary transcription factors HNF1β and SOX9[25,33], but inhibits TGFBR2[32,34]. Indeed, Gene Ontology (GO) term enrichment analysis of our RNA-seq data suggested active TGFβ signalling in hepatocyte-derived peripheral cholangiocytes, but not in normal peripheral cholangiocytes (Supplementary Table 1, down pC versus HpC). Moreover, we found high levels of phosphorylated SMAD3 (pSMAD3) in the nuclei of periportal HNF1-positive epithelial liver cells and in whole-liver nuclear extracts in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice at P60 (Extended Data Fig. 6b, c).

We functionally validated these findings by showing that the TGFβ inhibitor SB-431542 blocks biliary differentiation and morphogenesis of hepatocytes that lack RBPJ and HNF6 in vitro (Extended Data Fig. 6d–f). Moreover, at >P120, HpBDs were still absent or severely truncated, and liver fibrosis persisted in 14 out of 16 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ mice that, like *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice, were cholestatic at P34–P53 (Fig. 3c, d and Extended Data Fig. 6c, g).

These findings led us to investigate whether activating TGFβ signalling in hepatocytes enhances HpBD formation. We intravenously injected P19–P24 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice with an AAV serotype 8 vector expressing constitutively active TGFBR1 from the eukaryotic translation elongation factor 1α1 (*Eef1a1*) promoter (AAV8-Eef1a1-caTgfbr1). At P80–P100, 9 out of 11 treated mice had a mature hierarchical biliary network, whereas 9 out of 9 untreated mice still had an immature homogeneous biliary network[35] (Fig. 3e). Reflecting improved bile drainage, cholestasis and liver fibrosis resolved faster in treated mice (Fig. 3f, g). We ruled out the possibility that AAV8-Eef1a1-caTgfbr1 is fibrogenic in *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice (Extended Data Fig. 6h). These results identify TGFβ signalling as the driver of transdifferentiation and morphogenesis in HpBD formation.

To exclude the possibility that transdifferentiation is limited to immature hepatocytes, we investigated whether adult hepatocytes can form HpBDs. To bypass potential adaptive processes in development, we deleted *Rbpj* and *Hnf6* and activated GFP in hepatocytes of P75 *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice by co-injecting AAV8-Ttr-Cre and AAV8-Ttr-Flp and transplanted the cells 1 week later into P31 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Rag2*$^{-/-}$*Il2rg*$^{-/-}$ mice (Fig. 4a). After P120, we found donor hepatocyte-derived mature peripheral cholangiocytes in 5.2 ± 0.8% of the portal areas containing donor cells (Fig. 4b). We also transplanted NOTCH signalling-competent wild-type hepatocytes, which produced donor-derived HpBDs in 78.3 ± 13.9% of such portal

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

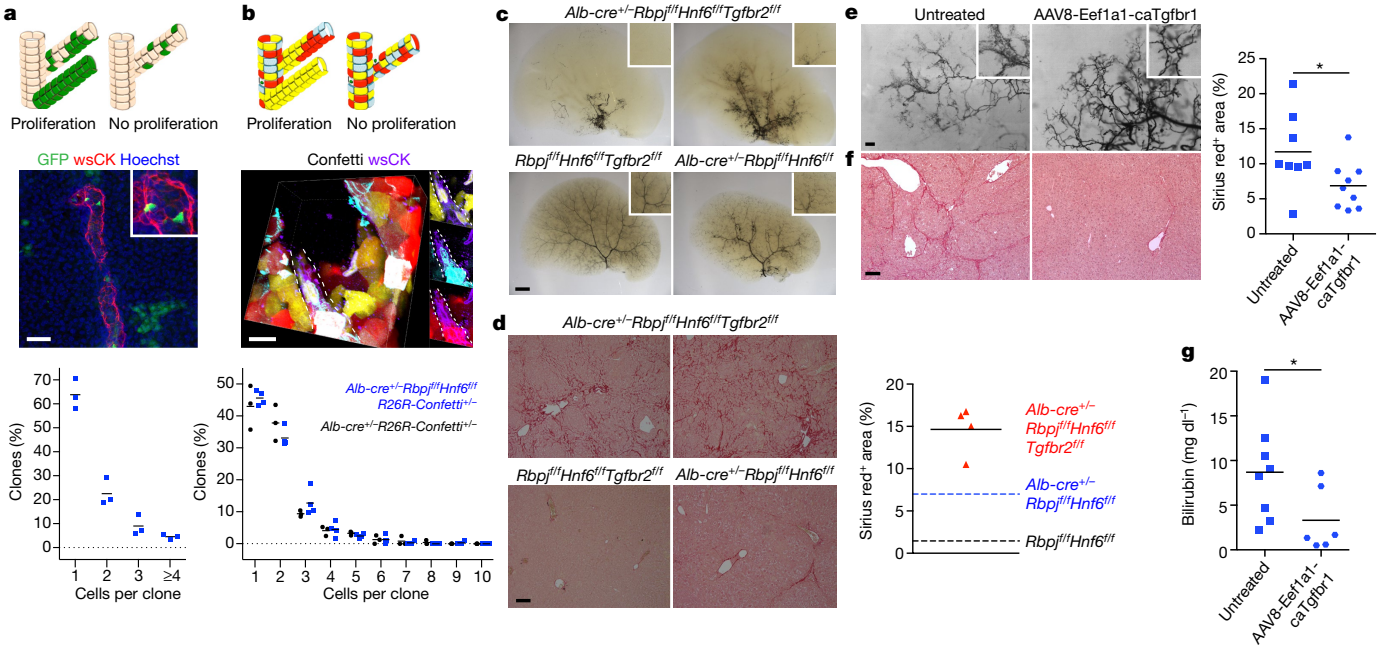


Fig. 3 | HpBD formation entails little proliferation and is driven by TGFβ signalling. a, Possible outcomes, maximum projection image and size distribution of clones in hepatocyte-fate-traced P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice ($n=3$). **b**, Possible outcomes, image stack volume projection and size distribution of clones in P150 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26R-Confetti*$^{+/-}$ ($n=4$) and *Alb-cre*$^{+/-}$*R26R-Confetti*$^{+/-}$ ($n=3$) mice. **c**, Ink visualization of biliary tree of >P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ ($n=16$), *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ ($n=4$) and *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ ($n=1$) mice. **d**, Sirius red staining with quantification in >P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ ($n=4$), *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ ($n=2$) and *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ ($n=2$) mice. Dotted lines represent the mean values of the indicated P120 mice from Extended Data Fig. 3f. **e**, **f**, Ink visualization of biliary tree and Sirius red staining with quantification in P100 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice that did ($n=9$) or did not ($n=8$) receive AAV8-Eef1a1-caTgfbr1. *$P=0.045$, two-sided Student's *t*-test. **g**, Serum total bilirubin in P36 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice that did ($n=6$) or did not ($n=8$) receive AAV8-Eef1a1-caTgfbr1. *$P=0.047$, two-sided Welch's *t*-test. Horizontal lines in **a**, **b**, **d**, **f** and **g** denote mean values. Scale bars, 2 mm (**c**), 500 μm (**e**), 100 μm (**d**, **f**), 50 μm (**a**), 20 μm (**b**).

areas because 31.6 ± 7.9% of the transplanted cells proliferated after transdifferentiation into cholangiocytes (Fig. 4a, c, d). These results show that adult hepatocytes, and transplanted hepatocytes, respond to the transdifferentiation-inducing and morphogenetic signals in the bile-duct-deficient liver and form HpBDs.

To determine whether our findings are relevant for human ALGS, we obtained liver samples from two patients who developed regenerative nodules containing pBDs[36] (Extended Data Table 1). The regenerative nodules contained CK7-positive pBDs, whereas nonregenerated liver tissue from the same patients showed abundant CK7-positive cells with hepatocyte morphology, indicative of metaplasia[14–17] (Fig. 4e). To determine whether TGFβ signalling is active in the new pBDs, we used pSMAD3 immunofluorescence. We found nuclear localization of pSMAD3 in 56.1 ± 6.1% of the pBDs in regenerative nodules, but not in pBDs in normal human liver (Fig. 4f). These results suggest that the TGFβ-mediated mechanism of HpBD formation identified in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice is also active in some patients with ALGS.

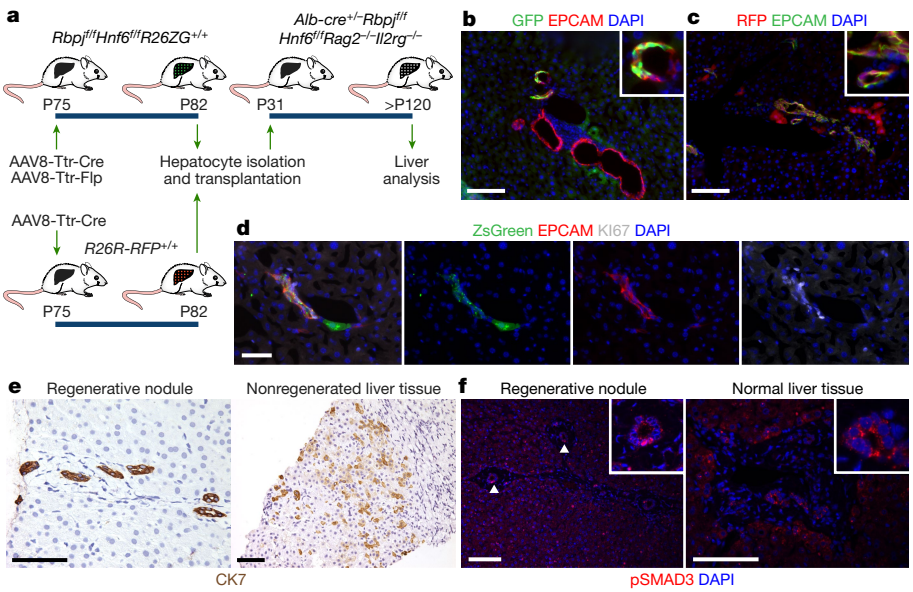


Fig. 4 | Clinical relevance and therapeutic potential of HpBD formation. **a**, Experimental design for hepatocyte transplantation. **b**, Immunofluorescence of P127 mouse ($n=5$) transplanted with adult GFP-expressing RBPJ- and HNF6-deficient hepatocytes. **c**, Immunofluorescence of P152 mouse ($n=4$) transplanted with adult RFP-expressing hepatocytes. **d**, Immunofluorescence of liver of P72 mouse ($n=2$) transplanted at P43 with hepatocytes isolated from P287 *Alb-cre*$^{+/-}$*R26R-ZsGreen*$^{+/+}$ mouse. **e**, **f**, Immunohistochemistry (**e**) and immunofluorescence (**f**) of ALGS ($n=2$) and normal ($n=1$) human livers. Arrowheads indicate nuclear pSMAD3 in pBDs. Scale bars, 100 μm (**b**, **c**, **e**, **f**), 50 μm (**d**).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

In conclusion, by showing that hepatocytes can convert into mature cholangiocytes and form a functional and stable biliary system, our findings establish that hepatocyte plasticity extends beyond metaplasia to transdifferentiation, and provide the first example, to our knowledge, of mammalian transdifferentiation building an organ structure de novo. Although hilar cholangiocytes are present in our mouse model, they fail to expand, resulting in severe cholestatic liver injury and strong pressure for hepatocytes to transdifferentiate into peripheral cholangiocytes. Analogously, cholangiocytes transdifferentiate into hepatocytes only when hepatocyte proliferation is completely suppressed[37]. The failure of hilar cholangiocytes to proliferate is probably caused by high TGFβ signalling in the cholestatic liver[38]. Accordingly, we identified TGFβ signalling as the driver of hepatocyte transdifferentiation and HpBD formation in our mouse model, and potentially also in patients with ALGS. Unlike for bile duct development[7–9], NOTCH signalling is not needed. Using clinically established AAV vectors and hepatocyte transplantation, we show that our findings are potentially translatable into therapies for ALGS and other diseases associated with a lack of bile ducts.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0075-5.

Received: 25 June 2016; Accepted: 22 March 2018;
Published online 2 May 2018.

1. Merrell, A. J. & Stanger, B. Z. Adult cell plasticity *in vivo*: de-differentiation and transdifferentiation are back in style. *Nat. Rev. Mol. Cell Biol.* **17**, 413–425 (2016).
2. Vanderpool, C. et al. Genetic interactions between hepatocyte nuclear factor-6 and Notch signaling regulate mouse intrahepatic bile duct development *in vivo*. *Hepatology* **55**, 233–243 (2012).
3. Tarlow, B. D. et al. Bipotential adult liver progenitors are derived from chronically injured mature hepatocytes. *Cell Stem Cell* **15**, 605–618 (2014).
4. Sekiya, S. & Suzuki, A. Hepatocytes, rather than cholangiocytes, can be the major source of primitive ductules in the chronically injured mouse liver. *Am. J. Pathol.* **184**, 1468–1478 (2014).
5. Kamimoto, K. et al. Heterogeneity and stochastic growth regulation of biliary epithelial cells dictate dynamic epithelial tissue remodeling. *eLife* **5**, e15034 (2016).
6. Tanimizu, N. et al. Progressive induction of hepatocyte progenitor cells in chronically injured liver. *Sci. Rep.* **7**, 39990 (2017).
7. Zong, Y. et al. Notch signaling controls liver development by regulating biliary differentiation. *Development* **136**, 1727–1739 (2009).
8. Sparks, E. E., Huppert, K. A., Brown, M. A., Washington, M. K. & Huppert, S. S. Notch signaling regulates formation of the three-dimensional architecture of intrahepatic bile ducts in mice. *Hepatology* **51**, 1391–1400 (2010).
9. Jeliazkova, P. et al. Canonical Notch2 signaling determines biliary cell fates of embryonic hepatoblasts and adult hepatocytes independent of Hes1. *Hepatology* **57**, 2469–2479 (2013).
10. Thorel, F. et al. Conversion of adult pancreatic α-cells to β-cells after extreme β-cell loss. *Nature* **464**, 1149–1154 (2010).
11. Stange, D. E. et al. Differentiated Troy$^+$ chief cells act as reserve stem cells to generate all lineages of the stomach epithelium. *Cell* **155**, 357–368 (2013).
12. Jain, R. et al. Plasticity of Hopx$^+$ type I alveolar cells to regenerate type II cells in the lung. *Nat. Commun.* **6**, 6727 (2015).
13. Tetteh, P. W. et al. Replacement of lost Lgr5-positive stem cells through plasticity of their enterocyte-lineage daughters. *Cell Stem Cell* **18**, 203–213 (2016).
14. Van Eyken, P., Sciot, R. & Desmet, V. J. A cytokeratin immunohistochemical study of cholestatic liver disease: evidence that hepatocytes can express 'bile duct-type' cytokeratins. *Histopathology* **15**, 125–135 (1989).
15. Libbrecht, L. et al. Peripheral bile duct paucity and cholestasis in the liver of a patient with Alagille syndrome: further evidence supporting a lack of postnatal bile duct branching and elongation. *Am. J. Surg. Pathol.* **29**, 820–826 (2005).
16. Ernst, L. M., Spinner, N. B., Piccoli, D. A., Mauger, J. & Russo, P. Interlobular bile duct loss in pediatric cholestatic disease is associated with aberrant cytokeratin 7 expression by hepatocytes. *Pediatr. Dev. Pathol.* **10**, 383–390 (2007).
17. Fabris, L. et al. Analysis of liver repair mechanisms in Alagille syndrome and biliary atresia reveals a role for notch signaling. *Am. J. Pathol.* **171**, 641–653 (2007).
18. Michalopoulos, G. K., Barua, L. & Bowen, W. C. Transdifferentiation of rat hepatocytes into biliary cells after bile duct ligation and toxic biliary injury. *Hepatology* **41**, 535–544 (2005).
19. Yanger, K. et al. Robust cellular reprogramming occurs spontaneously during liver regeneration. *Genes Dev.* **27**, 719–724 (2013).
20. Tanimizu, N., Nishikawa, Y., Ichinohe, N., Akiyama, H. & Mitaka, T. Sry HMG box protein 9-positive (Sox9$^+$) epithelial cell adhesion molecule-negative (EpCAM-) biphenotypic cells derived from hepatocytes are involved in mouse liver regeneration. *J. Biol. Chem.* **289**, 7589–7598 (2014).
21. Nagahama, Y. et al. Contributions of hepatocytes and bile ductular cells in ductular reactions and remodeling of the biliary system after chronic liver injury. *Am. J. Pathol.* **184**, 3001–3012 (2014).
22. Li, L. et al. Alagille syndrome is caused by mutations in human *Jagged1*, which encodes a ligand for Notch1. *Nat. Genet.* **16**, 243–251 (1997).
23. Oda, T. et al. Mutations in the human *Jagged1* gene are responsible for Alagille syndrome. *Nat. Genet.* **16**, 235–242 (1997).
24. McDaniell, R. et al. *NOTCH2* mutations cause Alagille syndrome, a heterogeneous disorder of the notch signaling pathway. *Am. J. Hum. Genet.* **79**, 169–173 (2006).
25. Poncy, A. et al. Transcription factors SOX4 and SOX9 cooperatively control development of bile ducts. *Dev. Biol.* **404**, 136–148 (2015).
26. Walter, T. J., Vanderpool, C., Cast, A. E. & Huppert, S. S. Intrahepatic bile duct regeneration in mice does not require Hnf6 or Notch signaling through Rbpj. *Am. J. Pathol.* **184**, 1479–1488 (2014).
27. Gong, A. Y. et al. Somatostatin stimulates ductal bile absorption and inhibits ductal bile secretion in mice via SSTR2 on cholangiocytes. *Am. J. Physiol. Cell Physiol.* **284**, C1205–C1214 (2003).
28. Li, B. et al. Adult mouse liver contains two distinct populations of cholangiocytes. *Stem Cell Reports* **9**, 478–489 (2017).
29. Peng, L. et al. RNA sequencing reveals dynamic changes of mRNA abundance of cytochromes P450 and their alternative transcripts during mouse liver development. *Drug Metab. Dispos.* **40**, 1198–1209 (2012).
30. Rezvani, M. et al. *In vivo* hepatic reprogramming of myofibroblasts with AAV vectors as a therapeutic strategy for liver fibrosis. *Cell Stem Cell* **18**, 809–816 (2016).
31. Fiorotto, R. et al. Notch signaling regulates tubular morphogenesis during repair from biliary damage in mice. *J. Hepatol.* **59**, 124–130 (2013).
32. Clotman, F. et al. Control of liver cell fate decision by a gradient of TGFβ signaling modulated by Onecut transcription factors. *Genes Dev.* **19**, 1849–1854 (2005).
33. Clotman, F. et al. The onecut transcription factor HNF6 is required for normal development of the biliary tract. *Development* **129**, 1819–1828 (2002).
34. Plumb-Rudewiez, N. et al. Transcription factor HNF-6/OC-1 inhibits the stimulation of the *HNF-3α/Foxa1* gene by TGF-β in mouse liver. *Hepatology* **40**, 1266–1274 (2004).
35. Tanimizu, N. et al. Intrahepatic bile ducts are developed through formation of homogeneous continuous luminal network and its dynamic rearrangement in mice. *Hepatology* **64**, 175–188 (2016).
36. Rougemont, A. L. et al. Bile ducts in regenerative liver nodules of Alagille patients are not the result of genetic mosaicism. *J. Pediatr. Gastroenterol. Nutr.* **61**, 91–93 (2015).
37. Raven, A. et al. Cholangiocytes act as facultative liver stem cells during impaired hepatocyte regeneration. *Nature* **547**, 350–354 (2017).
38. Mu, X. et al. Epithelial transforming growth factor-β signaling does not contribute to liver fibrosis but protects mice from cholangiocarcinoma. *Gastroenterology* **150**, 720–733 (2016).

Acknowledgements The authors received the following support: H.W.: NIH R01 DK107553, CIRM DISC1-08792, NIH P30 DK026743. S.S.H.: NIH R01 DK078640, NIH R01 DK107553, NIH P30 DK078392. J.R.S.: NIH T32 DK060414, A. P. Giannini Foundation. S.N.T.K. and B.Y.H.: NIH T32 GM008568. M.R.: Deutsche Forschungsgemeinschaft RE 3749/1-1. F.C.: NIH T32 DK060414, Jane Coffin Childs Memorial Fund. H.Y.L.: Eli and Edythe Broad Regeneration Medicine and Stem Cell Fellowship. The authors thank Donghui Wang (UCSF Preclinical Therapeutics Core), Chris Her (UCSF Liver Center Cell Biology Core), Vinh Nguyen (UCSF Flow Cytometry Core) and Matt Kofron and Mike Muntifering (CCHMC Nikon Center of Excellence Confocal Imaging Core) for technical support, Nick Timchenko, Maria Shrower and Abigail Roebker (CCHMC) for reagents and technical assistance, Rik Derynck (UCSF) for advice and Pamela Derish (UCSF) for manuscript editing.

Reviewer information *Nature* thanks S. Forbes, W. Goessling and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions H.W. and S.S.H. conceived the study. The authors contributed to designing and executing the experiments and data analysis as follows: J.R.S.: Figs. 1–4, Extended Data Figs. 1–6. K.A.H.: Figs. 1, 3, 4, Extended Data Figs. 3, 5, 6. S.N.T.K.: Figs. 1–4, Extended Data Fig. 1, 3–6, Extended Data Table 1. B.Y.H.: Fig. 4. A.E.C.: Extended Data Fig. 6. B.D.: Extended Data Fig. 6. R.A.K.: Fig. 2, Extended Data Fig. 4. F.C.: Extended Data Fig. 2. M.R.: Fig. 4. H.Y.L.: Fig. 3, Extended Data Fig. 3. A.N.M.: Fig. 4, Extended Data Table 1. A.-L.R.: Fig. 4, Extended Data Table 1. P.R.: Fig. 4, Extended Data Table 1. S.S.H.: Figs. 1–4, Extended Data Figs. 3, 5. H.W.: Fig. 2, Extended Data Fig. 4, Extended Data Table 1. H.W., S.S.H., J.R.S. and S.N.T.K. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0075-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0075-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.S.H. or H.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Mice. *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice[2,26] (mixed background), *R26R-RFP*$^{+/+}$ mice[39] (C57BL/6) and *R26NZG*$^{+/+}$ mice[40] (FVB) were previously reported. Flp-reporter mice were generated by crossing *R26NZG*$^{+/+}$ mice with *EIIa-cre*$^{+/+}$ mice[41] (C57BL/6) to remove the Cre-reporter element and then crossing out the *EIIa-cre*. These *R26ZG*$^{+/+}$ mice were crossed with *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice to generate *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice. *R26R-Confetti*$^{+/+}$ mice[42] (C57BL/6) were crossed with *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice to generate *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26R-Confetti*$^{+/-}$ mice. *Tgfbr2*$^{f/f}$ mice[43] (C57BL/6) were crossed with *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice to generate *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ mice. Because *Tgfbr2* (68.39 cM) and *Hnf6* (41.93 cM) are both on chromosome 9, recombinants were generated at 0.1356 (35 out of 258 mice) observed frequency (0.26 expected frequency). Different founder recombinants were intercrossed to generate *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ mice. *R26R-ZsGreen*$^{+/+}$ mice[44] (C57BL/6) were crossed with *Alb-cre*$^{+/-}$ mice to generate *Alb-cre*$^{+/-}$*R26R-ZsGreen*$^{+/+}$ mice. *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice were crossed with *Rag2*$^{-/-}$*Il2rg*$^{-/-}$ mice[45,46] (mixed background) to generate *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Rag2*$^{-/-}$*Il2rg*$^{-/-}$ mice. Male and female mice of the indicated age and genotype were chosen randomly for inclusion in experiments. The mouse used as a positive control for biliary gene expression was a C57BL/6 wild-type mouse fed choline-deficient diet (MP Biomedicals) and given 0.15% (w/v) ethionine (Sigma-Aldrich) in the drinking water (CDE diet) for 3 weeks. All mice were kept under barrier conditions. All procedures were approved by the Institutional Animal Care and Use Committee at UCSF or CCHMC.

Adeno-associated virus. The AAV-Ttr-Flp plasmid was generated by removing Cre from AAV-Ttr-Cre[47] and replacing it with Flpo[48] from pPGKFlpobpA (Addgene 13793); viruses were produced by Vector Biolabs and used at a high dose of 1×10^{12}–3×10^{12} vg or low dose of 4×10^{11} vg. The AAV-Eef1a1-caTgfbr1 plasmid was built by VectorBuilder to contain the activating T204D[49] mutation; virus was produced by Vector Biolabs and used at a dose of 1×10^{11} vg. Titres were determined using qPCR. Viruses were delivered by tail vein injection in volumes $\leq$100 µl to prevent hydrodynamic effects.

Tissue collection. Livers from *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$, *R26R-Confetti*$^{+/-}$ and transplanted mice were perfused with ice-cold PBS followed by 4% paraformaldehyde (PFA). Samples were cut into slices and fixed overnight in 4% PFA at 4 °C. Livers from *R26R-RFP*$^{+/+}$, *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$, *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ and *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice were fixed overnight in 4% PFA at 4 °C. For thin sectioning, samples were moved to 30% sucrose overnight at 4 °C to cryopreserve and then embedded in OCT (Tissue-Tek) or processed for paraffin embedding and sectioning. A Leica CM3050 S cryostat was used to cut 6 µm cryosections for staining and imaging. For 3D and sparse-labelling clonal analysis, liver samples were embedded in 4% agarose and sectioned on a Leica VT100 S vibratome. For clonal analysis in *R26R-Confetti*$^{+/-}$ mice, liver samples were cut into 1 mm slivers, immunostained, washed in PBS, equilibrated in 30% sucrose and embedded in OCT. The 1 mm slivers were then cryosectioned at 100 µm, stained with Hoechst and mounted in PBS for confocal imaging. For analysis of Flp-based hepatocyte fate tracing in *R26ZG*$^{+/+}$ mice, liver samples were fixed in neutral-buffered formalin containing zinc (Z-Fix, Anatech), embedded in paraffin and sectioned to 5 µm.

Human liver tissue. Explant samples from the regenerative nodule and nonregenerated liver tissue of a 3-year-old male patient with ALGS were previously described; regenerative nodule and nonregenerated liver tissue contained the same heterozygous *JAG1* exon 1–26 deletion[36]. Samples were obtained with patient consent and approval from the Commission Cantonale d'Ethique de la Recherché CCER. Biopsy samples from the regenerative nodule and nonregenerated liver tissue of a 15-year-old male patient with ALGS caused by a heterozygous c.499 T > A (p.W167R) *JAG1* mutation and resection samples from a histologically normal region of the liver of a 35-year-old male undergoing surgery for metastasis of rectal adenocarcinoma were obtained with patient consent and approval from the UCSF Institutional Review Board.

Immunostaining and histology. Cryosections were blocked in 10% normal serum and permeabilized in 0.3% Triton-X before staining with primary and secondary antibodies listed in Supplementary Tables 2 and 3. Paraffin-embedded samples were deparaffinized and underwent antigen retrieval in sodium citrate buffer (Bio-Genex or Vector Labs) or Tris EDTA buffer (10 mM Tris Base, 1 mM EDTA, 0.05% Tween 20, pH 9.0) in a steamer or pressure cooker for 15 min before blocking and permeabilization. Thin sections were mounted in FluorSave (MilliporeSigma) or 50% glycerol. Vibratome sections were stained free-floating in 12-well dishes and cleared in Focus Clear (Cedarlane Laboratories) before being mounted in Mount Clear (Cedarlane Laboratories) for confocal imaging. All samples stained with anti-wsCK antibody underwent antigen retrieval in 100 mM Tris buffer (pH 9.5) at 55 °C for 2 h before blocking. Staining with anti-pSMAD3 antibody used a biotinylated secondary antibody followed by avidin/biotin-based peroxidase and tyramide amplification. Staining with anti-SSTR2 antibody used an HRP secondary antibody followed by tyramide-based amplification. M.O.M. Kit (Vector Laboratories) was used for antibodies raised in mice. DAPI or Hoechst was used to stain nuclear DNA. Chromogenic detection used a horseradish peroxidase (HRP) secondary antibody followed by avidin/biotin-based peroxidase amplification and DAB substrate exposure. Sirius red staining was performed on deparaffinized and rehydrated samples using a 0.1% Picrosirius red solution (Dudley Corporation and Newcomer Supply) and 0.5% acetic acid water washes or was done at Peninsula Histopathology Laboratory.

Imaging. Thin sections were imaged on an Olympus BX51 upright microscope and cultured cells were imaged on an Olympus IX71 inverted microscope using Openlab software (PerkinElmer). Chromogenic stains were imaged on a Leica DM 1000 LED. Fiji[50] and/or Photoshop (Adobe) were used to process (brightness, contrast and gamma) and merge channels. Thick sections for 3D analysis of connectivity and clonal analysis were imaged on a Leica upright AOBS confocal microscope and processed and analysed using Imaris (Bitplane) or Volocity (PerkinElmer) software. For *R26R-Confetti*$^{+/-}$ mouse and pSMAD3 analysis, images were acquired using a Nikon A1R GaAsP inverted SP confocal microscope and NIS elements software and processed and analysed using Imaris software. Sirius red-stained sections were imaged using a Cytation 5 cell imaging multi-mode reader (BioTek).

Ink injection. A catheter was inserted in a retrograde fashion into the common bile duct of post-mortem mice and waterproof ink (Higgins) was slowly injected. Left liver lobes were dehydrated in 1:1 methanol:water followed by 100% methanol. Ink was visualized by tissue clearing in 1:2 benzyl alcohol:benzyl benzoate (BABB) solution and imaged on a Leica M205A or Nikon SMZ800 stereoscope.

Cholangiocyte isolation. Nonparenchymal liver cells were isolated from >P115 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ and *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice as previously described[51]. Cells were resuspended at 1×10^7 cells ml^{-1} in DMEM/2% FBS and blocked with Mouse Fc Block (BD Biosciences) for 30 min. Cells were incubated with fluorochrome-conjugated antibodies (Supplementary Table 2) and DBA-FITC (Vector Laboratories) for 30 min, washed with cold Dulbecco's PBS (DPBS) three times and resuspended in DMEM/2% FBS. Sytox Red (Thermo Fisher Scientific) was added to label dead cells before sorting. Unstained and single-stained cells were used for compensation. Specificity of DBA binding was verified with a GalNAc (Sigma)-blocked control as previously described[52]. Cells were analysed and sorted on a FACSAria III using FACSDiva software (BD Biosciences). From the CD11b$^-$ CD31$^-$CD45$^-$ population, EPCAM$^+$DBA$^-$ cells were collected as peripheral cholangiocytes and EPCAM$^+$DBA$^+$ as hilar cholangiocytes. FlowJo (FlowJo, LLC) was used to analyse data and generate charts. Cells were either sorted into DMEM/2% FBS, pelleted and snap frozen, or sorted directly into extraction buffer for RNA purification.

Hepatocyte isolation. Hepatocytes were isolated from >P115 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice by two-step collagenase (Worthington) perfusion followed by purification through a Percoll gradient. Cells were resuspended at 1×10^6 cells per 100 µl in Hanks buffer with 10% FBS and incubated with OC2-2F8 antibody (Supplementary Table 2) for 1 h on ice. Cells were washed with cold DPBS twice and resuspended in Hanks/10% FBS. Fluorochrome-conjugated secondary antibody (Supplementary Table 3) was added and cells were incubated for 30 min on ice followed by two washes with cold DPBS. Cells were blocked with 5% normal rat serum (Jackson Immuno) in Hanks buffer for 10 min on ice. Cells were then incubated with fluorochrome-conjugated antibodies (Supplementary Table 2) for 30 min on ice. Cells were then washed in cold DPBS three times and resuspended in Williams E medium/2% FBS. Sytox Red (Thermo Fisher Scientific) was added to label dead cells before sorting. Unstained and single-stained cells were used for compensation. Cells were analysed and sorted on a FACSAria III using FACSDiva software (BD Biosciences). From the CD11b$^-$CD31$^-$CD45$^-$EPCAM$^-$ population, OC2-2F8$^+$ cells were collected as hepatocytes (Supplementary Fig. 2). Cells were either sorted into DMEM/2% FBS, pelleted and snap frozen, or sorted directly into extraction buffer for RNA purification.

Hepatocyte transplantation. Hepatocytes were isolated from donor mice by two-step collagenase perfusion followed by purification through a Percoll gradient. 1×10^6 viable cells were resuspended in 80–100 µl of Williams E medium with glutamine. Transplantation was performed by transdermal intrasplenic injection of the cell suspension under isoflurane anaesthesia.

DDC diet feeding. Mice received PicoLab Mouse Diet 20, 5058 (LabDiet) with 0.1% 3,5-diethoxycarbonyl-1,4-dihydrocollidine (DDC; Sigma-Aldrich) for the indicated durations.

qPCR and gene expression analysis. Genomic DNA was isolated from cells using QIAamp DNA Micro Kit or DNeasy Blood & Tissue Kit (Qiagen). RNA was extracted using Trizol Reagent (Thermo Fisher Scientific) and purified by isopropanol precipitation (cells in collagen gel) or purified using RNAeasy Mini Kit (Qiagen) (liver tissue). Reverse transcription was performed using qScript cDNA Supermix (Quanta Biosciences). qPCR was performed using SYBR green reagent (Thermo Fisher Scientific) in a ViiA 7 system (Thermo Fisher Scientific). Reactions were performed in triplicate, and expression was normalized to an *Rbpj*

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

(genotyping) or *Gapdh* (gene expression) reference and quantified using the $\Delta\Delta C_t$ method. Primers are listed in Supplementary Table 4.

RNA-seq. RNA was purified from FACS-isolated cells using PicoPure Kit (Thermo Fisher Scientific). RNA quality was assessed using RNA 6000 Pico Kit on a 2100 Bioanalyzer (Agilent). Samples with an RNA integrity number (RIN) ≥ 7.7 and at least 15 ng of RNA were used to construct sequencing libraries using Clontech Low Input Library Prep Kit v2. Libraries were sequenced on a HiSeq 3000, 10 samples per lane, with single-end 50 bp reads. Raw reads were aligned to the mm10 mouse genome with annotations provided by UCSC using CobWEB, a proprietary Burrows-Wheeler Transform method. Reads per kilobase per million (RPKM) were calculated from aligned reads using the expectation-maximization algorithm. RPKM was thresholded at 1, $\log_2$ transformed, normalized using the DESeq algorithm and baselined to the median of all samples. Analyses were performed on transcripts with RPKM > 5 in all samples of at least 1 experimental condition ($n = 17{,}793$ transcripts). These reasonably expressed transcripts were used in principle component analysis. All transcripts with fold change > 3 in at least 1 of the 3 possible pairwise comparisons ($n = 6{,}464$) were selected, and a one-way ANOVA was performed to identify significantly differential genes with FDR-corrected $P < 0.05$ ($n = 4{,}997$). Venn diagrams were used to identify unique and shared gene signatures. Gene sets were submitted to https://ToppGene.cchmc.org for identification of pathway and biological process enrichments.

Western blot. Nuclear and cytoplasmic extracts were generated from whole liver using previously described buffers with protease and phosphatase inhibitors[53]. Samples were run on SDS–PAGE 4–20% Tris-glycine gradient gels, electrophoretically transferred to nitrocellulose membrane and probed with antibodies. Signals were detected by ECL western blotting substrate (GE Healthcare). The membrane was stripped and reprobed with anti-actin antibody to verify and normalize protein loading using densitometry. Quantification was performed using ImageJ software.

Serum chemistry. Blood was collected from post-mortem *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice and controls of ages P20 to $\geq$P150 and tested for serum total bilirubin (TecoDiagnostics) and alkaline phosphatase, alanine aminotransferase and aspartate aminotransferase (ADVIA XPT clinical chemistry system, Siemens). For serial bilirubin measurements, blood was collected by retro-orbital venipuncture and tested for total bilirubin every 2 weeks from P60 to P189 in four different litters. Blood was collected from *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ mice and controls by retro-orbital venipuncture and tested for serum total bilirubin (TecoDiagnostics). Blood was collected from *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice intravenously injected with AAV8-Eef1a1-caTgfbr1 and controls by retro-orbital venipuncture and serum total bilirubin was measured as previously reported[54] using a Synergy 2 microplate reader (BioTek). Serum absorbance at 540 nm was subtracted from serum absorbance at 450 nm. A linear trendline equation of serum absorbance versus serum total bilirubin was determined by independently measuring absorbance and total bilirubin (TecoDiagnostics) of known cholestatic and noncholestatic serum samples. This equation was used to convert absorbance readings to serum total bilirubin levels.

In vitro conversion assay. In vitro 3D culture of hepatocytes to induce biliary conversion was carried out as previously reported[21] with the following modifications. Hepatocytes were isolated from 8–10-week-old *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice injected with 1×10^{12} vg of AAV8-Ttr-Cre 2 weeks before to delete *Rbpj* and *Hnf6*. Hepatocytes were isolated by two-step collagenase perfusion followed by purification through a Percoll gradient, all in the absence of serum. Cells were grown on Primaria plates for 6 days to form spheroids. Spheroids were cultured in collagen gels (Cultrex Rat Collagen I, Lower Viscosity, Trevigen) in the presence of 10 μM SB-431542 (Selleckchem) or vehicle (DMSO). The medium was changed every other day.

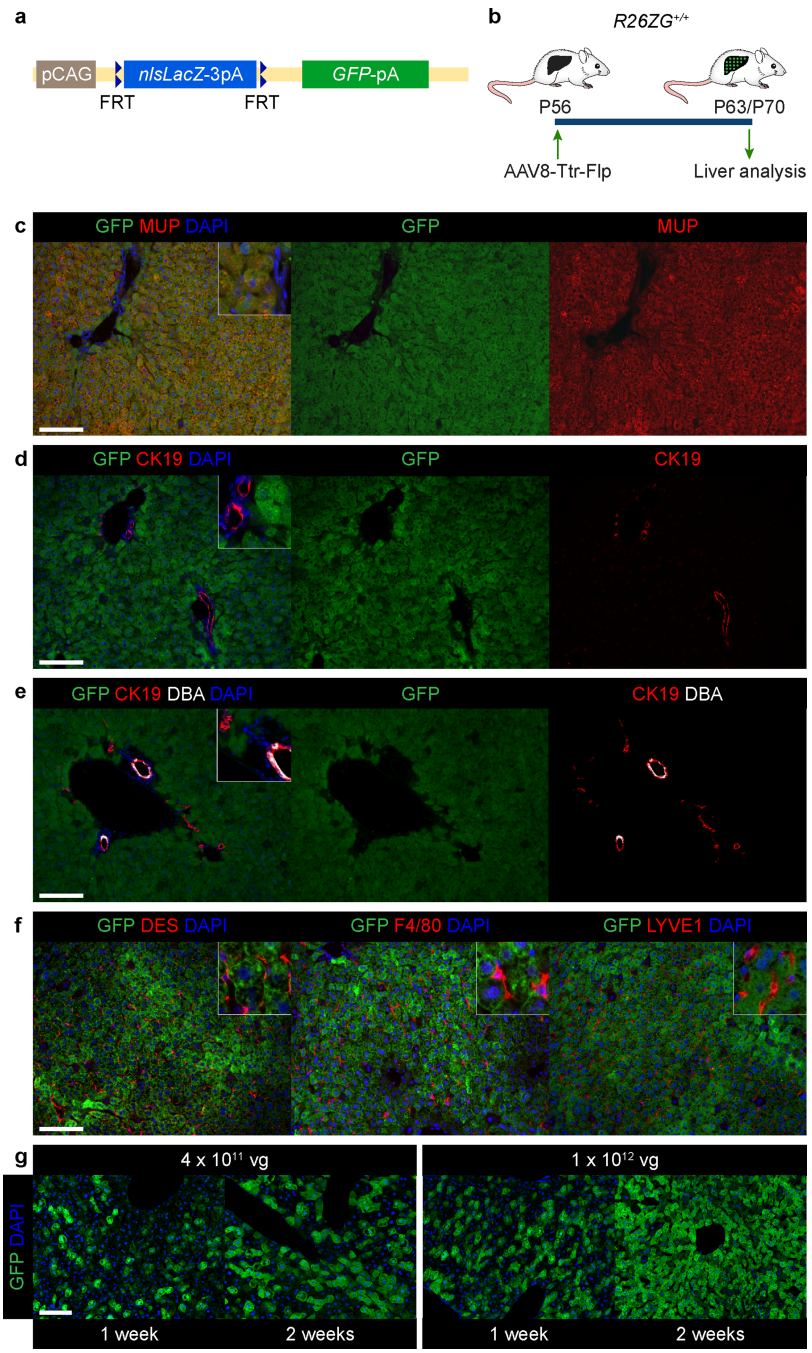
Quantification and statistics. For sparse-labelling clonal analysis, eight liver regions from each of three *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice were analysed. GFP-positive cells in all clones within 100 μm-thick sections were manually counted using Imaris software for visualization. GFP-positive cells in direct contact were considered clones. Clones that extended to and potentially beyond the *x*, *y* or *z* boundaries were excluded. For clonal analysis in *R26R-Confetti*$^{+/-}$ mice, 30–40 μm *z*-stack images of wsCK-positive DBA-negative pBDs and wsCK-positive DBA-positive hBDs were visualized in 3D with Imaris software. The module 'surfaces' within Imaris was used to render a 3D surface created on an intensity value on a per channel basis. The rendered surface per channel was masked to the Hoechst nuclear stain. 3D pBD masks were used to manually count cells per clone. Multiple portal regions were analysed for clones at P90 (14 and 17 per *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26R-Confetti*$^{+/-}$ and 8 and 9 per *Alb-cre*$^{+/-}$*R26R-Confetti*$^{+/-}$ mouse) and P150 (7–10 per *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26R-Confetti*$^{+/-}$ and 8–10 per *Alb-cre*$^{+/-}$*R26R-Confetti*$^{+/-}$ mouse). To determine labelling efficiencies in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice, for each mouse, 20 random fields were analysed for wsCK-positive DBA-negative peripheral cholangiocytes (~350 cells) and 5 random fields were analysed for hepatocytes (~1,500 cells). For proliferation analysis, OPN-positive cells in four random portal fields from each of four mice were analysed for KI67 and CK19 staining. For hepatocyte-fate-tracing analysis in DDC diet-fed mice, ten random portal fields from four *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ and three *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice were analysed. For quantification in transplantation experiments, all portal areas with donor-derived cells in a section from at least three lobes per mouse were examined and scored for the presence or absence of donor hepatocyte-derived EPCAM-positive cholangiocytes. For human samples, 63 bile ducts between the 2 patient samples and 45 bile ducts in the control sample were scored for the presence of nuclear pSMAD3. For quantification of Sirius red staining, liver samples were stained in batches and sections of whole lobes were imaged at equal exposure. Using Fiji software, a threshold was set for Sirius red-positive area within each lobe and the percentage of the total area that was Sirius red positive was measured and is reported. Charts were generated in Prism 6 or 7 (GraphPad). Researchers were not blinded when analysing results. $P < 0.05$ was considered statistically significant. Experiments were replicated independently once (Fig. 2a, b, d–g, 3a, b, d, g, Extended Data Figs. 1c–g, 2c, e, 3a–g, 4b–d, 5a, c, d, f, 6c, g; Supplementary Table 1), at least twice (Fig. 1d, e, 2c, 3f, 4b–f, Extended Data Figs. 5b, e, 6a, b, e, f, h; Supplementary Video 1) or at least three times (Fig. 1b, c, 3c, e, Extended Data Figs. 2b, 4a; Supplementary Video 2). No statistical methods were used to predetermine sample size.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The Gene Expression Omnibus (GEO) accession number for the RNA-seq data is GSE108315. Additional data are available from the corresponding authors on reasonable request.

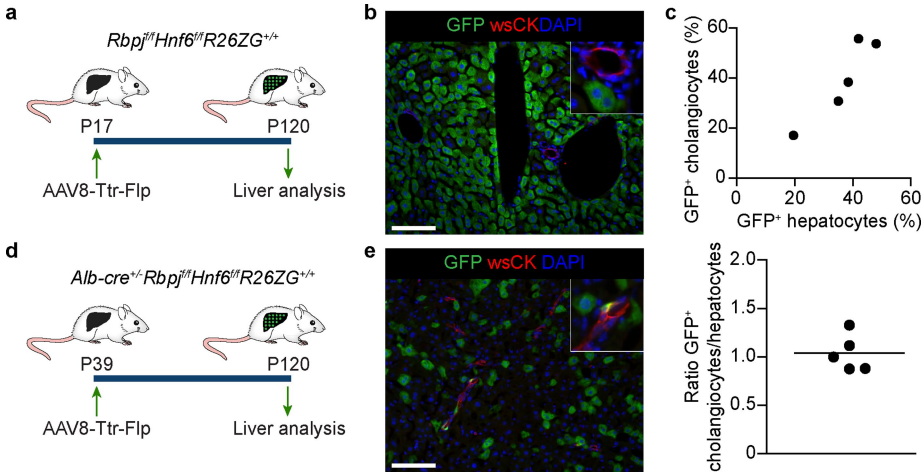
39. Luche, H., Weber, O., Nageswara Rao, T., Blum, C. & Fehling, H. J. Faithful activation of an extra-bright red fluorescent protein in "knock-in" Cre-reporter mice ideally suited for lineage tracing studies. *Eur. J. Immunol.* **37**, 43–53 (2007).
40. Yamamoto, M. et al. A multifunctional reporter mouse line for Cre- and FLP-dependent lineage analysis. *Genesis* **47**, 107–114 (2009).
41. Lakso, M. et al. Efficient in vivo manipulation of mouse genomic sequences at the zygote stage. *Proc. Natl Acad. Sci. USA* **93**, 5860–5865 (1996).
42. Snippert, H. J. et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
43. Levéen, P. et al. Induced disruption of the transforming growth factor beta type II receptor gene in mice causes a lethal inflammatory disorder that is transplantable. *Blood* **100**, 560–568 (2002).
44. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).
45. Shinkai, Y. et al. RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. *Cell* **68**, 855–867 (1992).
46. Cao, X. et al. Defective lymphoid development in mice lacking expression of the common cytokine receptor γ chain. *Immunity* **2**, 223–238 (1995).
47. Malato, Y. et al. Fate tracing of mature hepatocytes in mouse liver homeostasis and regeneration. *J. Clin. Invest.* **121**, 4850–4860 (2011).
48. Raymond, C. S. & Soriano, P. High-efficiency FLP and PhiC31 site-specific recombination in mammalian cells. *PLoS One* **2**, e162 (2007).
49. Wieser, R., Wrana, J. L. & Massagué, J. GS domain mutations that constitutively activate TβR-I, the downstream signaling component in the TGF-β receptor complex. *EMBO J.* **14**, 2199–2208 (1995).
50. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
51. Dorrell, C. et al. Prospective isolation of a bipotential clonogenic liver progenitor cell in adult mice. *Genes Dev.* **25**, 1193–1203 (2011).
52. Chen, Z., Wang, M., Xiang, Q., Sun, Z. & Zhang, R. The lectin of *Dolichos biflorus* agglutinin recognises glycan epitopes on the surface of a subset of cardiac progenitor cells. *Cell Biol. Int.* **37**, 1238–1245 (2013).
53. Orellana, D. et al. Calmodulin controls liver proliferation via interactions with C/EBPβ-LAP and C/EBPβ-LIP. *J. Biol. Chem.* **285**, 23444–23456 (2010).
54. Siggs, O. M., Schnabl, B., Webb, B. & Beutler, B. X-linked cholestasis in mouse due to mutations of the P4-ATPase ATP11C. *Proc. Natl Acad. Sci. USA* **108**, 7890–7895 (2011).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



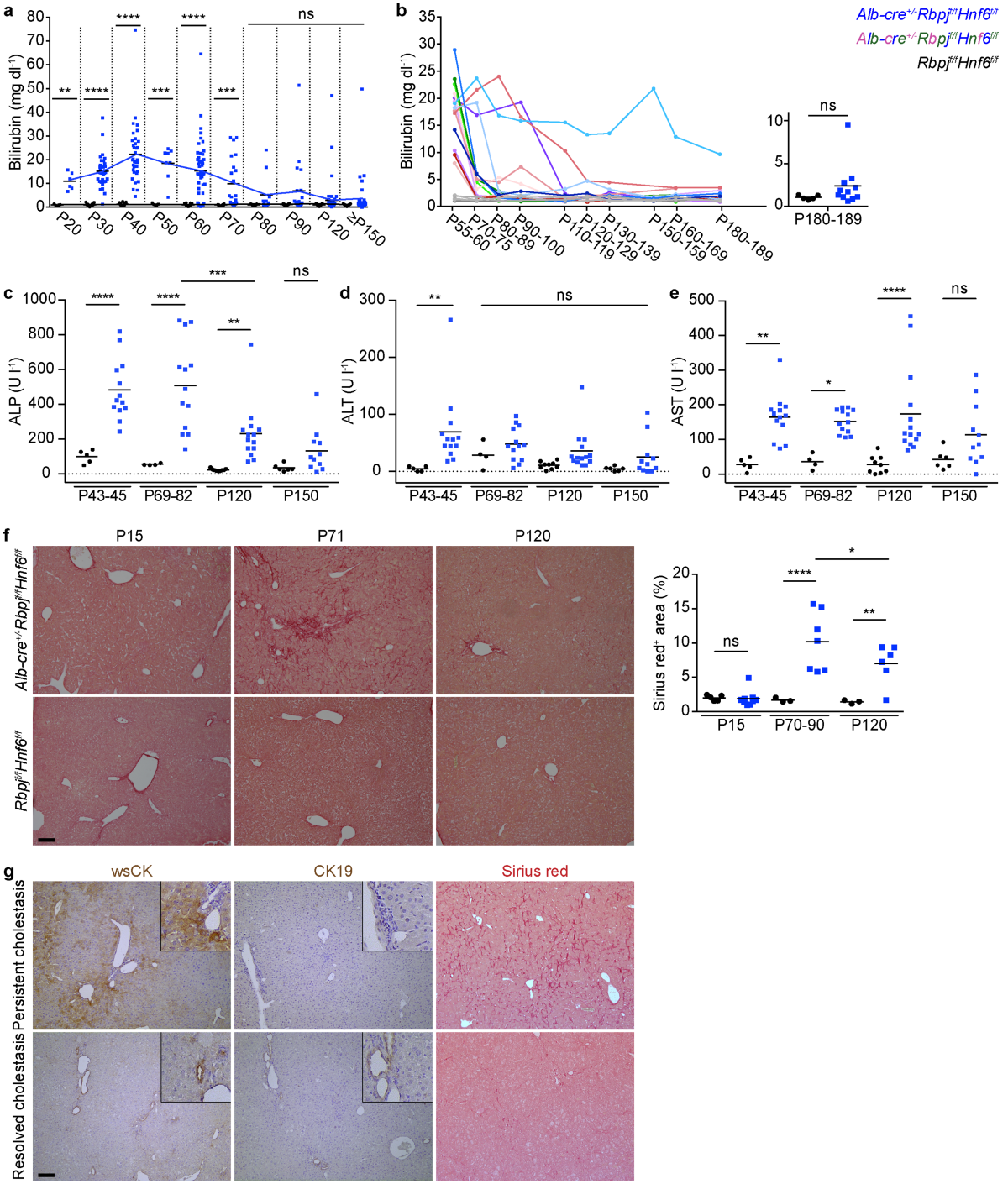
Extended Data Fig. 1 | Flp-based hepatocyte fate tracing. a, *R26ZG* allele. **b**, Experimental design for establishing efficient, specific and constitutive labelling of hepatocytes in normal adult $R26ZG^{+/+}$ mice. **c**–**f**, Immunofluorescence of $R26ZG^{+/+}$ mouse liver ($n = 2$) for GFP and the hepatocyte marker major urinary protein (MUP) (**c**), peripheral and hilar cholangiocyte marker CK19 (**d**, **e**), hilar cholangiocyte marker DBA (**e**), hepatic stellate cell marker desmin (DES), macrophage marker F4/80 and endothelial cell marker LYVE1 (**f**) 2 weeks after intravenous injection of 1×10^{12} viral genomes (vg) of AAV8-Ttr-Flp. **g**, Reporter activation in $R26ZG^{+/+}$ mice 1 and 2 weeks after intravenous injection of the indicated dose of AAV8-Ttr-Flp ($n = 1$ for each dose and time point). Scale bars, 100 μm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



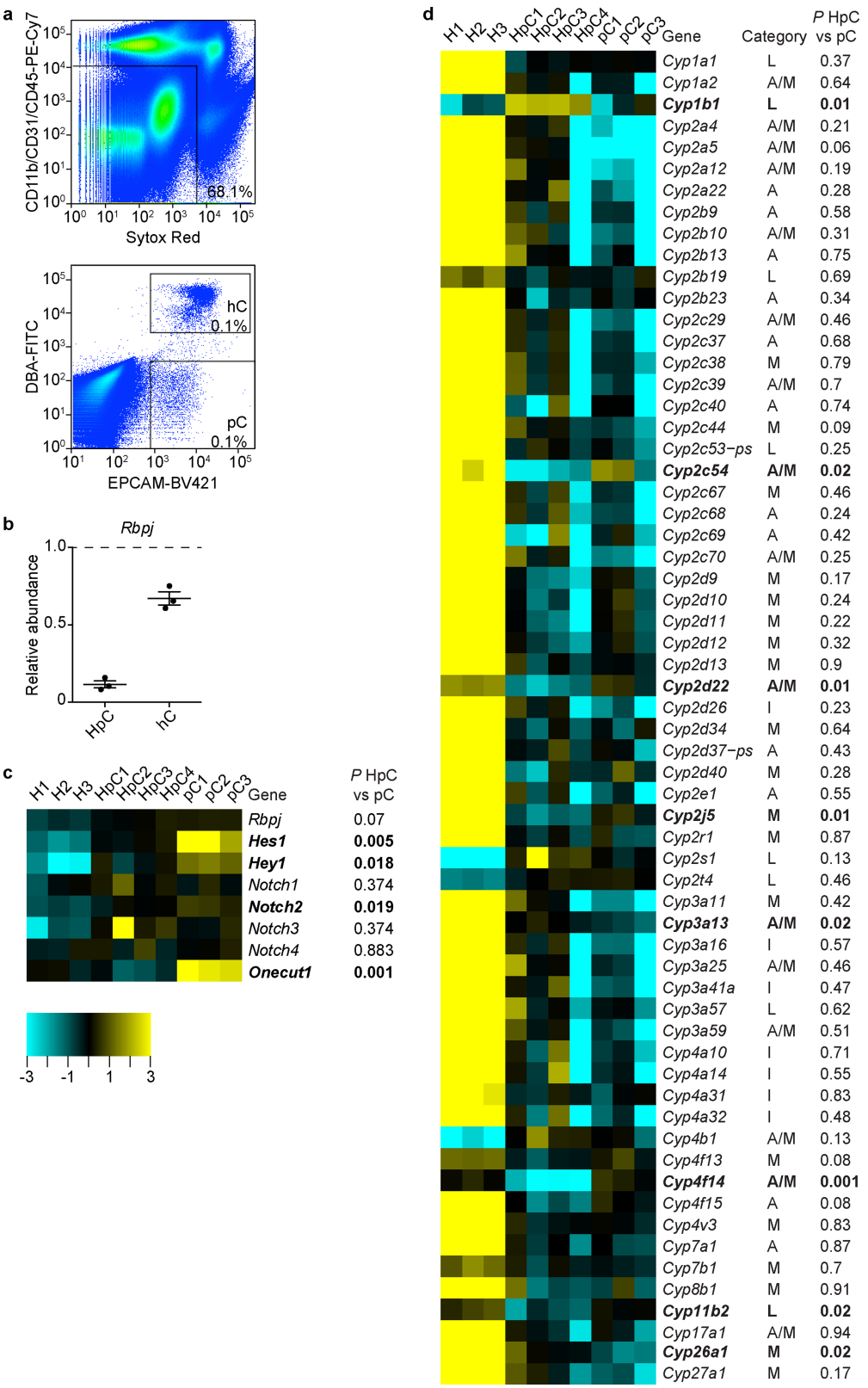
Extended Data Fig. 2 | Efficiency of hepatocyte fate tracing in mice born with or without pBDs. a, **b**, Experimental design for hepatocyte fate tracing at P17 and immunofluorescence at P120 in *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice ($n = 4$). **c**, Correlation of GFP labelling efficiency between hepatocytes and peripheral cholangiocytes in hepatocyte-fate-traced P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice ($n = 5$; top and bottom). Horizontal line denotes the mean value. **d**, **e**, Experimental design for hepatocyte fate tracing at P39 and immunofluorescence at P120 in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mice ($n = 3$). Scale bars, 100 μm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



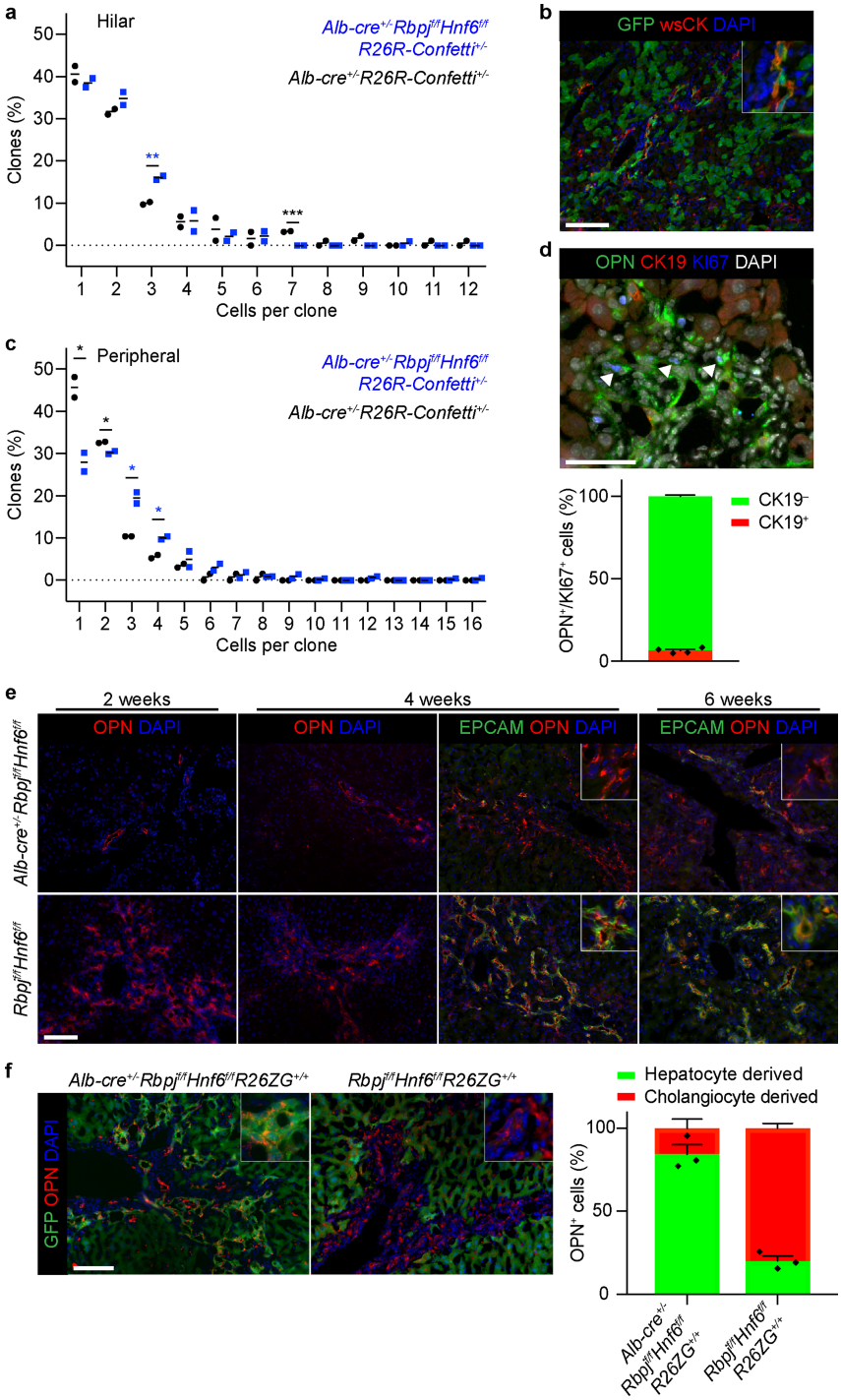
Extended Data Fig. 3 | HpBDs relieve cholestasis and liver injury. **a**, Serum total bilirubin levels in P20–P29 ($n=6$), P30–P39 ($n=33$), P40–P49 ($n=35$), P50–P59 ($n=8$), P60–P69 ($n=46$), P70–P79 ($n=22$), P80–P89 ($n=13$), P90–P119 ($n=20$), P120–P149 ($n=52$) and ≥P150 ($n=40$) *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ and P20–P29 ($n=5$), P30–P39 ($n=24$), P40–P49 ($n=19$), P50–P59 ($n=7$), P60–P69 ($n=27$), P70–P79 ($n=10$), P80–P89 ($n=11$), P90–P119 ($n=10$), P120–P149 ($n=41$) and ≥P150 ($n=25$) *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice. $^{**}P=0.0011$ at P20–P29, $^{****}P=1.7\times10^{-13}$ at P30–P39, $^{****}P=8.2\times10^{-12}$ at P40–P49, $^{***}P=0.00019$ at P50–P59, $^{****}P=8.4\times10^{-10}$ at P60–P69, $^{***}P=0.00055$ at P70–P79, not significant (ns; $P=0.090$) at P80–P89, not significant ($P=0.050$) at P90–P119, not significant ($P=0.052$) at P120–P149 and not significant ($P=0.064$) at ≥P150, two-sided Welch's *t*-test. **b**, Serial measurements of serum total bilirubin levels in *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ ($n=14$) and *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ ($n=5$) mice. $P=0.11$, two-sided Welch's *t*-test. **c**–**e**, Serum alkaline phosphatase (ALP), alanine aminotransferase (ALT) and aspartate aminotransferase (AST) levels in P43–P45 ($n=13$), P69–P82 ($n=13$), P120 ($n=14$) and P150 ($n=11$) *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ and P43–P45 ($n=5$), P69–P82 ($n=4$), P120 ($n=9$) and P150 ($n=6$) *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice. In **c**: $^{****}P=0.000058$ at P43–P45, $^{****}P=0.000019$ at P69–P82, $^{**}P=0.0064$ at P120, not significant ($P=0.22$) at P150 and $^{***}P=0.00010$ at P120 versus P69–P82, two-way ANOVA followed by Holm–Sidak multiple comparison test. In **d**: $^{**}P=0.0050$ at P43–P45, not significant ($P=0.47$) at P69–P82, not significant ($P=0.30$) at P120 and not significant ($P=0.47$) at P150, two-way ANOVA followed by Holm–Sidak multiple comparison test. In **e**: $^{**}P=0.0024$ at P43–P45, $^{*}P=0.015$ at P69–P82, $^{****}P=0.000073$ at P120 and not significant ($P=0.061$) at P150, two-way ANOVA followed by Holm–Sidak multiple comparison test. **f**, Sirius red staining in P15 ($n=9$), P70–P90 ($n=7$) and P120 ($n=6$) *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ and P15 ($n=5$), P70–P90 ($n=3$) and P120 ($n=3$) *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice with quantification. Not significant ($P=0.94$) at P15, $^{****}P=0.000095$ at P70–P90, $^{**}P=0.0074$ at P120 and $^{*}P=0.027$ at P120 versus P70–90, two-way ANOVA followed by Holm–Sidak multiple comparison test. **g**, Immunohistochemistry and Sirius red staining in P313 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice with persistent or resolved cholestasis ($n=1$ each). Horizontal lines in **a**–**f** denote mean values. Scale bars, 100 μm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



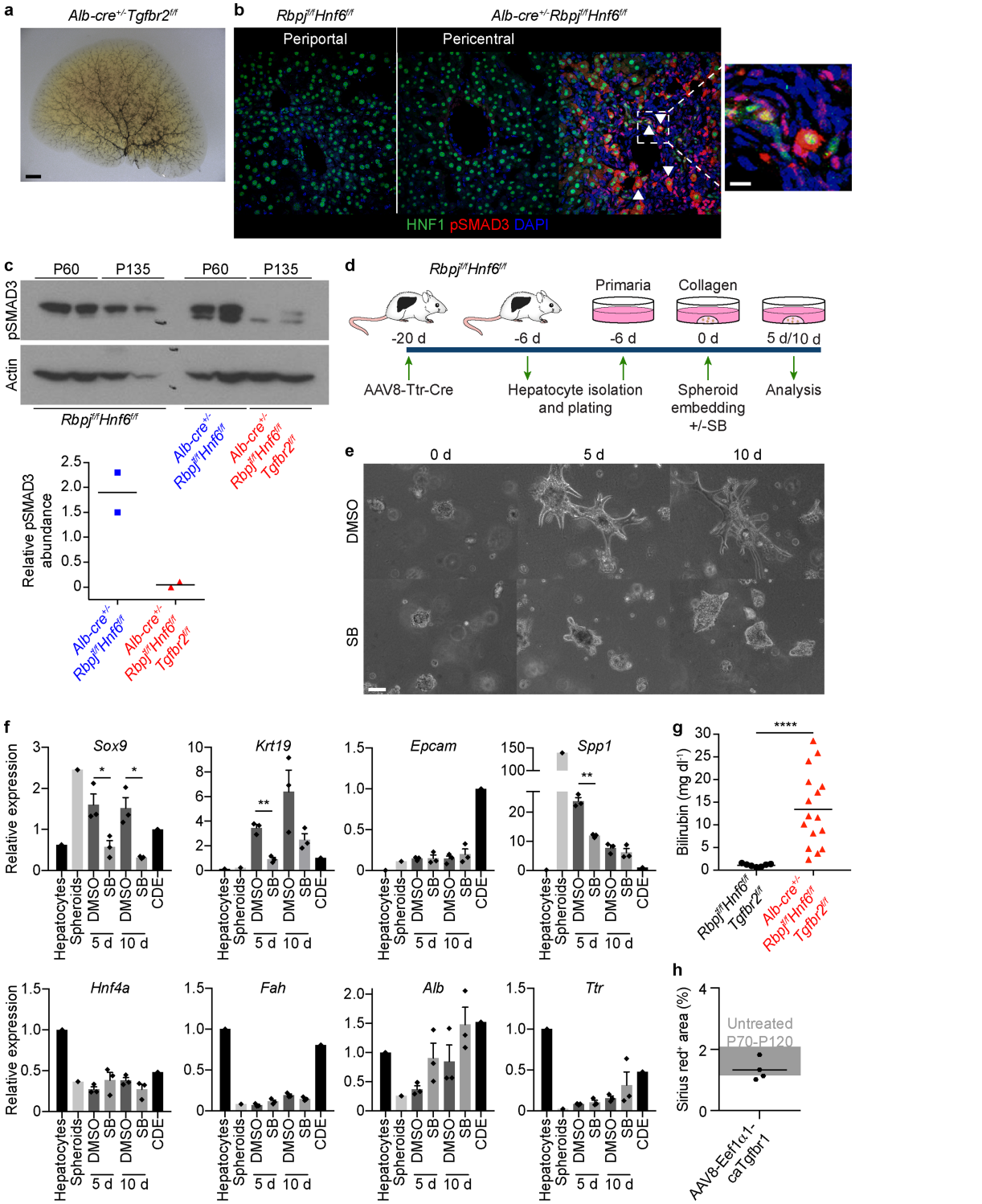
Extended Data Fig. 4 | Isolation and gene expression profiling of hepatocyte-derived peripheral cholangiocytes. a, FACS gates for peripheral cholangiocyte (pC; $EPCAM^{+}DBA^{-}$) and hilar cholangiocyte (hC; $EPCAM^{+}DBA^{+}$) isolation from *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ and *Rbpj*$^{f/f}$ *Hnf6*$^{f/f}$ mice. **b**, qPCR analysis of floxed *Rbpj* (*Rbpj*$^{f/f}$) genomic DNA in hepatocyte-derived pC (HpC) and hC isolated from *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$ *Hnf6*$^{f/f}$ mice relative to hepatocytes isolated from *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice (dashed line; $n = 3$ each). Data were normalized to a downstream genomic region of *Rbpj* to control for gene copy number. Data are mean ± s.e.m. **c**, **d**, RNA-seq analysis of normal pC ($n = 3$ mice), HpC ($n = 4$ mice) and RBPJ- and HNF6-deficient hepatocytes (H; $n = 3$ mice). **c**, Heat map of genes reflecting deletion of *Rbpj* and *Hnf6* (also known as *Onecut1*). *Rbpj* mRNA is present in this knockout mouse as a truncated transcript that does not produce a functional protein[26]. **d**, Heat map of all differentially expressed CYP genes distinguishing genes associated with mature (M), adolescent (A) and immature (I) hepatocyte differentiation or low expression in the liver (L)[29]. $P < 0.05$, one-way ANOVA, FDR-corrected; fold change > 3 (**c**, **d**, except *Rbpj* and *Notch1*–*Notch4*). Bold genes denote $P < 0.05$ for HpC versus pC, two-sided Student's *t*-test (**c**, **d**).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 5 | Proliferation in HpBDs and reactive ductules. **a**, Size distribution of wsCK-positive DBA-positive hilar cholangiocyte clones in P90 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26R-Confetti*$^{+/-}$ ($n=2$) and *Alb-cre*$^{+/-}$*R26R-Confetti*$^{+/-}$ ($n=2$) mouse livers. $^{**}P=0.0079$ for three cells and $^{***}P=0.00092$ for seven cells, two-sided Student's *t*-test. **b**, Immunofluorescence of reactive ductules in hepatocyte-fate-traced P32 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ mouse liver ($n=3$). **c**, Size distribution of wsCK-positive DBA-negative peripheral cholangiocyte clones in P90 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26R-Confetti*$^{+/-}$ ($n=2$) and *Alb-cre*$^{+/-}$*R26R-Confetti*$^{+/-}$ ($n=2$) mouse livers. $^{*}P=0.032$ for one cell, $^{*}P=0.024$ for two cells, $^{*}P=0.020$ for three cells and $^{*}P=0.014$ for four cells, two-sided Student's *t*-test. Horizontal lines in **a** and **c** denote mean values. **d**, Immunofluorescence and breakdown of OPN-positive KI67-positive cells based on CK19 expression in P54 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mouse liver ($n=4$). Arrowheads indicate OPN-positive KI67-positive CK19-negative cells. **e**, Immunofluorescence of liver of >P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ and *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice after DDC diet feeding for 2 ($n=1$ each), 4 ($n=3$ each) and 6 ($n=1$ each) weeks. **f**, Immunofluorescence of liver and breakdown of OPN-positive cells based on hepatocyte fate tracing in >P120 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ ($n=4$) and *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*R26ZG*$^{+/+}$ ($n=3$) mice fed DDC diet for 5 weeks starting 1 week after hepatocyte fate tracing was induced. Data in **d** and **f** are mean ± s.e.m. Scale bars, 100 μm (**b**, **e**, **f**), 50 μm (**d**).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | TGFβ signalling in hepatocyte transdifferentiation. a, Ink visualization of biliary tree of P32 *Alb-cre*$^{+/-}$*Tgfbr2*$^{f/f}$ mouse ($n = 2$). **b**, Immunofluorescence of P60 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ and *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mouse livers ($n = 2$ each). Arrowheads indicate pSMAD3-positive HNF1-positive nuclei. **c**, Western blot with quantification of pSMAD3 in nuclear extracts from *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$, *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ and *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ mouse livers ($n = 2$ each). Source data are shown in Supplementary Fig. 1. **d**–**f**, Experimental design (**d**) and results of analysis of the effect of TGFβ signalling on biliary differentiation of adult RBPJ- and HNF6-deficient hepatocytes cultured in 3D for the indicated number of days (d). **e**, Phase-contrast images of RPBJ- and HNF6-deficient hepatocyte spheroids embedded in collagen gels and cultured in the presence or absence of the TGFβ inhibitor SB-431542 (SB) for the indicated number of days. **f**, Relative expression levels of cholangiocyte and hepatocyte genes in freshly isolated hepatocytes and spheroids before and after embedding in collagen gels. Gene expression in the liver of a mouse fed choline-deficient ethionine-supplemented (CDE) diet was used as a positive control. Data are from three independent cultures per treatment in a representative experiment ($n = 2$). *$P = 0.038$ for *Sox9* 5 days, *$P = 0.044$ for *Sox9* 10 days, **$P = 0.0034$ for *Krt19* 5 days and **$P = 0.0071$ for *Spp1* 5 days, two-sided Welch's *t*-test. **g**, Serum total bilirubin levels in P34–P53 *Alb-cre*$^{+/-}$*Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ ($n = 16$) and *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$*Tgfbr2*$^{f/f}$ ($n = 7$) mice. ****$P = 0.000024$, two-sided Welch's *t*-test. **h**, Quantification of Sirius red staining in P58–P100 *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice after intravenous injection of AAV8-Eef1a1-caTgfbr1 at P20 ($n = 4$). Grey area represents the range of liver collagen in the indicated *Rbpj*$^{f/f}$*Hnf6*$^{f/f}$ mice from Extended Data Fig. 3f. Horizontal lines in **c**, **g** and **h** denote mean values. Data in **f** are mean ± s.e.m. Scale bars, 2 mm (**a**), 100 μm (**e**), 50 μm (**b**) and 10 μm (**b**, inset).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Characterization of human subjects and samples

Human subject	Gender	ALGS genotype	ALGS phenotype	Age at liver analysis	Liver samples
ALGS patient	Male	Heterozygous *JAG1* exon 1-26 deletion	Neonatal cholestasis, bile duct paucity, liver fibrosis, portal hypertension, hepatocellular carcinoma, liver transplantation; left pulmonary artery stenosis; butterfly vertebrae; facial features	3 years	Regenerative nodule and nonregenerated tissue (segment 5 of right lobe and segment 8 of right lobe of explanted liver, respectively)
ALGS patient	Male	Heterozygous *JAG1* exon 4 mutation (c.499T>A; p.W167R)	Failure to thrive; hepatomegaly, bile duct paucity, cholestasis, liver fibrosis; left pulmonary artery stenosis; facial features	15 years	Regenerative nodule and nonregenerated tissue (biopsies of segment 3 of left liver lobe and segment 5 of right liver lobe, respectively)
Normal control	Male	Normal *JAG1* and *NOTCH2* sequences	None	35 years	Histologically normal tissue (resection of segment 4 of left liver lobe for metastasis of rectal adenocarcinoma)

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0086-2

Male-killing toxin in a bacterial symbiont of *Drosophila*

Toshiyuki Harumoto[1]* & Bruno Lemaitre[1]*

Several lineages of symbiotic bacteria in insects selfishly manipulate host reproduction to spread in a population[1], often by distorting host sex ratios. *Spiroplasma poulsonii*[2,3] is a helical and motile, Gram-positive symbiotic bacterium that resides in a wide range of *Drosophila* species[4]. A notable feature of *S. poulsonii* is male killing, whereby the sons of infected female hosts are selectively killed during development[1,2]. Although male killing caused by *S. poulsonii* has been studied since the 1950s, its underlying mechanism is unknown. Here we identify an *S. poulsonii* protein, designated Spaid, whose expression induces male killing. Overexpression of Spaid in *D. melanogaster* kills males but not females, and induces massive apoptosis and neural defects, recapitulating the pathology observed in *S. poulsonii*-infected male embryos[5–11]. Our data suggest that Spaid targets the dosage compensation machinery on the male X chromosome to mediate its effects. Spaid contains ankyrin repeats and a deubiquitinase domain, which are required for its subcellular localization and activity. Moreover, we found a laboratory mutant strain of *S. poulsonii* with reduced male-killing ability and a large deletion in the *spaid* locus. Our study has uncovered a bacterial protein that affects host cellular machinery in a sex-specific way, which is likely to be the long-searched-for factor responsible for *S. poulsonii*-induced male killing.

Endosymbiotic bacteria have evolved sophisticated strategies to manipulate their hosts to increase their transmission, and sex ratio distorters of arthropods are examples of this manipulation. These bacteria are transmitted exclusively through female hosts, and as a result, several lineages have evolved the ability to bias infections towards females, either by turning males into females (feminization), causing clonal reproduction (parthenogenesis), or eliminating males (male killing)[1]. Male killing has independently evolved in at least six bacterial taxa, including *Spiroplasma*, *Wolbachia*, and *Rickettsia*[1]. The genetic basis of male killing is a longstanding mystery in the field of insect symbiosis. Male killing by *S. poulsonii* (Fig. 1a) was first described in the 1950s in *Drosophila*[2]. Previous studies of *S. poulsonii* attributed the selective killing of male progeny to an unknown substance called 'androcidin', assumed to be secreted by the bacterium[12]. The identification of this toxin has been hampered by the lack of practical methods for molecular biology, as has been the case for other symbiotic bacteria.

S. poulsonii symbionts of *D. melanogaster* (strain MSRO for '*melanogaster* sex ratio organism') kill all male progeny (for example, MSRO-H99; Extended Data Fig. 1a). We unexpectedly identified a *S. poulsonii* mutant strain that shows reduced male-killing ability (MSRO-SE; the partial male-killing strain), where almost half of the male progeny survived (Extended Data Fig. 1a–c). The reduced male killing was not linked to host genetic background or low bacterial titre (Extended Data Fig. 1b, d). To identify the genetic basis of reduced male killing, we sequenced the genome of MSRO-SE and compared it with that of MSRO-H99 (Extended Data Fig. 2). We found a candidate gene that was altered in the partial male-killing strain, encoding a 1,065-amino-acid protein with ankyrin repeats and the OTU (ovarian tumour) deubiquitinase domain. We named this *S. poulsonii* protein Spaid (*S. poulsonii* androcidin; Fig. 1b) based on our subsequent functional studies. We focused our analysis on this gene, because (i) both ankyrin repeats and the OTU domain are conserved in proteins across eukaryotes and are also present in some

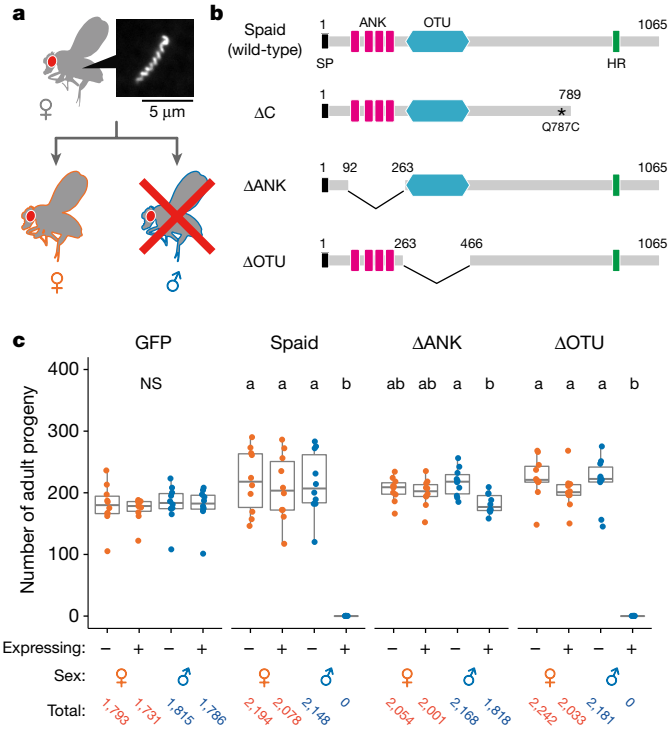


Fig. 1 | Expression of Spaid selectively eliminates male offspring. **a**, Male killing in *Drosophila* induced by *S. poulsonii*. Infected females (top) produce only female offspring (bottom). Inset shows a male-killing *S. poulsonii* of *D. melanogaster* detected by DNA staining. **b**, Protein structure of Spaid, which contains ankyrin repeats (ANK, red), the OTU deubiquitinase domain (blue), an N-terminal signal peptide (SP, black), and a C-terminal hydrophobic region (HR, green). Spaid ΔC of the partial male-killing strain encodes a protein with an amino acid substitution (Q787C) and a C-terminal truncation. The structures of the two deletion constructs of Spaid (ΔANK and ΔOTU) are also indicated. The numbers represent amino acid residues. **c**, The number of adult progeny obtained from crosses between the *actin–GAL4* line and four *UAS* transgenic lines (GFP, Spaid, ΔANK, and ΔOTU; $n = 10$ independent crosses for each transgene). The *UAS–GFP* line was used as a negative control. We counted the number of resultant offspring (females, red; males, blue) having both *actin–GAL4* and *UAS* (+) and siblings having only *UAS* as internal controls (−). Different lowercase letters indicate statistically significant differences (Steel–Dwass test (see Supplementary Table 2)). $P < 0.0001$ in all instances except ΔANK ($P < 0.05$). NS, not significant ($P > 0.05$). Box plots indicate the median (bold line), the 25th and 75th percentiles (box edges), and the range (whiskers). Dot plots show all data points individually. The total numbers of adult counts for each genotype and sex are shown at the bottom.

[1]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. *e-mail: toshiyuki.harumoto@epfl.ch; bruno.lemaitre@epfl.ch

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

bacterial effectors that manipulate host cellular processes[13,14], and (ii) *spaid* was located on a putative plasmid[15] (Extended Data Fig. 2b), like other bacterial virulence factors. Further analysis predicted an N-terminal signal peptide for secretion and a C-terminal hydrophobic region (Fig. 1b). The *spaid* locus in the partial male-killing strain contained an 828-bp deletion (Extended Data Fig. 3), resulting in a truncated protein lacking the hydrophobic region, as well as a single amino acid substitution (Q787C) (ΔC; Fig. 1b). Of note, this gene was not present in an earlier published version of the *S. poulsonii* MSRO genome[16] (Supplementary Data), and we found no obvious homologous proteins in our BLAST searches.

To test whether Spaid is responsible for male killing, we used the GAL4/UAS system to overexpress the gene as a C-terminal green fluorescent protein (GFP) fusion protein in *D. melanogaster*. Notably, Spaid expression with the *actin–GAL4* ubiquitous driver eliminated all male offspring, but had no impact on female emergence (Fig. 1c). Thus, Spaid kills flies in a sex-dependent manner.

Male killing in *Drosophila* induced by *S. poulsonii* is associated with abnormal apoptosis[7,9] and neural disorganization[5,8,9] during embryogenesis. The mechanism of the neural defects is not known, but has been suggested to be independent of apoptosis[8,9]. If Spaid is the bona fide male-killing factor, its expression in embryos should phenocopy the pathology described above. We employed the *nanos-GAL4* maternal driver to express Spaid during early embryogenesis and monitored apoptosis (Fig. 2; with terminal deoxynucleotidyl transferase dUTP nick-end labelling (TUNEL) staining) and neural organization (Extended Data Fig. 4; using the neuron-specific marker Elav[17]). The sex of embryos was determined by antibody staining for the Sex lethal (Sxl) protein, which is expressed only in females[18]. We found that Spaid expression induced strong apoptosis and neural disorganization in male but not female embryos (Fig. 2 and Extended Data Fig. 4). The level of apoptosis in male embryos that expressed Spaid increased as their development proceeded (Fig. 2c), similar to embryos infected with *S. poulsonii*[9]. Although numerous cells underwent apoptosis in male embryos that expressed Spaid, the neural cells did not appear to suffer ectopic cell death (Extended Data Fig. 4c). Thus, the expression

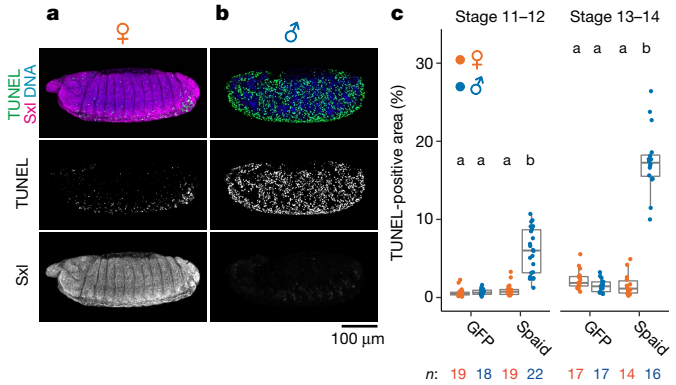


Fig. 2 | Expression of Spaid reproduces male-killing phenotypes during embryogenesis. a, **b**, Representative images of stage 13–14 female (**a**, $n = 14$) and male (**b**, $n = 16$) embryos maternally expressing Spaid, stained for apoptosis (TUNEL; green), Sxl (magenta), and DNA (blue). Single-channel images of TUNEL and Sxl are also shown. **c**, Quantification of TUNEL signals in stage 11–12 and 13–14 embryos (females, red; males, blue). Different characters indicate statistically significant differences ($P < 0.0001$; Steel–Dwass test; see Supplementary Table 2). Box and dot plots as in Fig. 1c. Sample sizes (*n*, number of embryos) are shown at the bottom. Embryos were co-stained for Elav, TUNEL, Sxl, and DNA, and selected channels are shown in **a**, **b** and Extended Data Fig. 4.

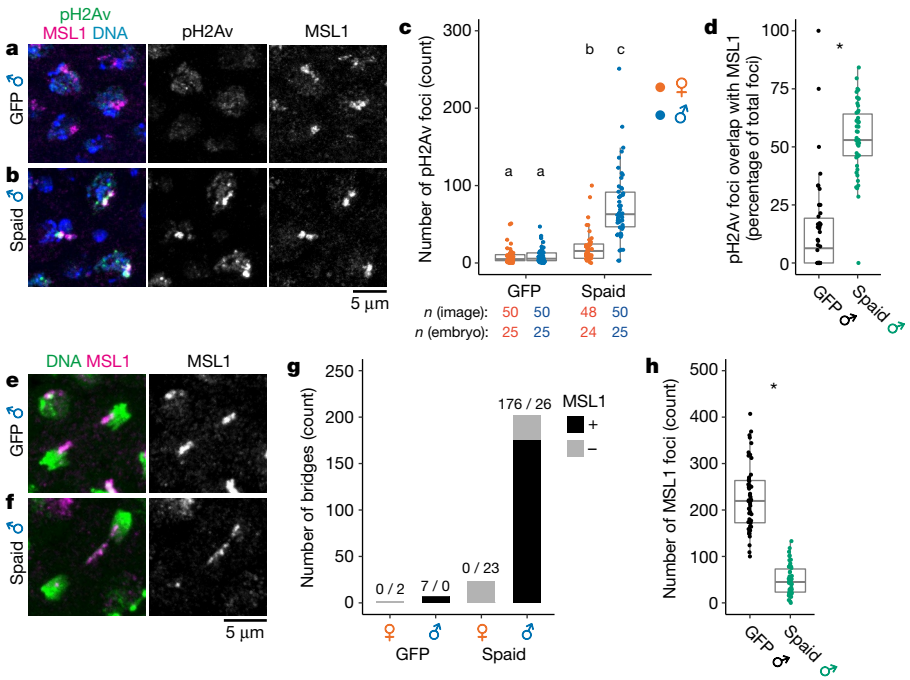


Fig. 3 | Spaid acts through the MSL complex. a, **b**, Epithelial cells in stage 9 male embryos expressing GFP (**a**) and Spaid (**b**), stained for pH2Av (green), MSL1 (magenta), and DNA (blue). **c**, Quantification of pH2Av foci in embryos expressing GFP and Spaid. **d**, Percentage of pH2Av foci overlapping with MSL1 signals in male embryos expressing; GFP, 6.3% (0–19.3%) (median (interquartile range)) and Spaid, 52.9% (46.2–64.2%). **e**, **f**, Dividing cells in stage 9 male embryos expressing GFP (**e**, proper segregation) and Spaid (**f**, a broken bridge) stained for DNA (green) and MSL1 (magenta). **g**, The number of chromatin bridges containing (black; numbers on the left) or not containing (grey; numbers on the right) MSL1 signals in embryos expressing GFP and Spaid. **h**, The number of MSL1 focal signals in male embryos expressing GFP (black) and Spaid (green). The same datasets of stage 8–10 embryos were analysed in **a**–**h** ($n = 50$ or 48 images per condition). Different lowercase letters or asterisks indicate significant differences: **c**, $P < 0.001$, Steel–Dwass test; **d**, $P < 0.0001$, χ^2 test; and **h**, $P < 0.0001$, two-tailed Mann–Whitney *U* test (see Supplementary Table 2). Box and dot plots in **c**, **d**, **h** are as in Fig. 1c and sample sizes are shown at the bottom in **c**. All *UAS* transgenes were expressed maternally.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

of Spaid is sufficient to induce the two different pathologies caused by male-killing *S. poulsonii*.

In *Drosophila* (XX female, XY male), the single male X chromosome is hyper-transcribed by twofold to equalize gene expression levels between sexes. This dosage compensation system is mediated by a protein–RNA complex called the male-specific lethal (MSL) complex, which is selectively recruited to the male X chromosome[19]. Previous studies have uncovered a link between the male-killing action of *S. poulsonii* and the host's dosage compensation machinery[6,10,11]. Genetic experiments revealed that *S. poulsonii* fails to kill males lacking the MSL components, while it can induce death in females ectopically expressing the MSL complex[6,10,11]. This suggests that *S. poulsonii* either targets the MSL complex directly or targets its downstream chromatin modifications (for example, acetylation of histone H4 at K16)[19]. Remarkably, we found that the expression of Spaid triggered massive apoptosis in transgenic female embryos engineered to express the MSL complex (Extended Data Fig. 5a–d), indicating that this toxin mediates its effects through the dosage compensation machinery. We previously showed that *S. poulsonii* infection triggers DNA damage and segregation defects on the male X chromosome, thereby activating male-specific apoptosis[11]. To test whether Spaid expression can reproduce these phenotypes, we monitored DNA damage on the male X chromosome by antibody staining for the phosphorylated histone H2Av (pH2Av)[20] and MSL1, a component of the MSL complex labelling the male X chromosome. In control embryos expressing GFP and female embryos expressing Spaid, a few H2Av foci were detected, whereas male embryos expressing Spaid exhibited numerous bright foci (Fig. 3a–c) that frequently overlapped with MSL1 signals (Fig. 3d). We also found that Spaid expression in male embryos caused the formation of chromatin bridges, of which 87.1% contained MSL1 signals (Fig. 3e–g). The MSL1-labelled chromosomes were often fragmented and unevenly distributed after mitosis (Fig. 3f; 47.7% of the MSL1-containing bridges). In addition, MSL1 signals were reduced in male embryos that expressed Spaid, and we observed this phenotype at an even earlier time point (Fig. 3h and Extended Data Fig. 5e, f; stages 8–10) than has been described for embryos infected by *S. poulsonii* (from stage 13 onward)[10]. This is probably because the high expression of Spaid driven by the GAL4/UAS system results in much more severe effects than *S. poulsonii* infection. Finally, we examined the subcellular distribution of Spaid in larval salivary glands, whose large cellular sizes facilitate detailed cytological analyses. In both sexes, Spaid–GFP was found at plasma membranes, in the cytoplasm, and throughout nuclei, but was absent from the nucleolus (Fig. 4a, b). Notably, Spaid–GFP was enriched on MSL1-labelled chromosomes in male nuclei (Fig. 4b). This result is consistent with its reliance on the dosage compensation machinery, although equivalent localization analyses in embryos are needed.

To better characterize the action of Spaid, we made deletion constructs lacking either ankyrin repeats (ΔANK) or the OTU domain (ΔOTU) (Fig. 1b). Expression of ΔANK with *actin–GAL4* did not kill males, pointing to a pivotal role of ankyrin repeats in Spaid activity (Fig. 1c). The expression of ΔOTU with *actin–GAL4* still eliminated adult males, but they were killed later in development (pupal stage) compared with wild-type Spaid (second instar larval stage) (Fig. 1c). Furthermore, and in contrast to wild-type Spaid, weaker expression of ΔOTU with *armadillo–GAL4* failed to kill males (Extended Data Fig. 6), indicating that the OTU domain is required for full activity. Notably, although its distribution pattern was otherwise indistinguishable from that of Spaid–GFP in salivary gland cells, ΔANK–GFP was not enriched on MSL-labelled chromosomes (Fig. 4c, d). By contrast, ΔOTU–GFP was only weakly detected within nuclei regardless of sex, and colocalization with MSL was not apparent (Fig. 4e, f). The male-killing activity of ΔOTU when overexpressed may be caused by excess amounts of protein overriding its localization defect. All of these observations can be reconciled in a model of the chromosomal targeting of Spaid in which the OTU domain promotes nuclear localization and ankyrin repeats interact with the MSL complex or downstream histone modifications (Extended Data Fig. 7).

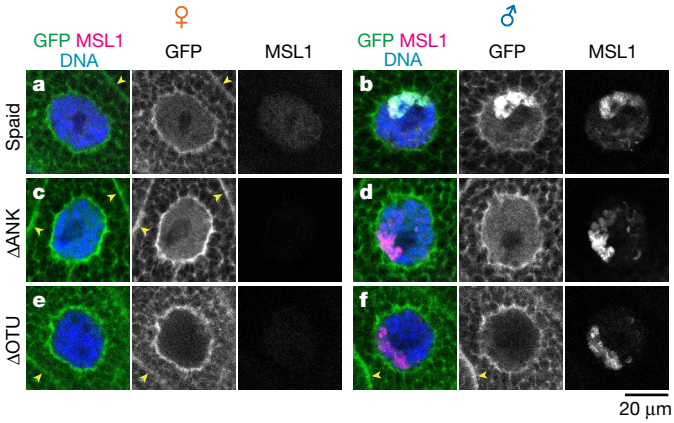


Fig. 4 | Subcellular localization of Spaid. a–f, Larval salivary glands expressing Spaid–GFP (**a**, female, $n = 13$; **b**, male, $n = 17$), ΔANK–GFP (**c**, female, $n = 9$; **d**, male, $n = 16$), and ΔOTU–GFP (**e**, female, $n = 15$; **f**, male, $n = 12$) stained for MSL1 (magenta) and DNA (blue). For GFP (green), raw fluorescence signals were detected. Magnified views of nuclei are shown. Dark spots inside nuclei in GFP images represent the nucleolus. Arrowheads indicate GFP signals associated with plasma membranes.

In summary, we have discovered an *S. poulsonii* gene that is likely to be responsible for male killing. Notably, expression of this single gene was sufficient to recapitulate the phenotypes associated with male killing. Our analysis revealed that Spaid is the sole ankyrin-repeat protein in the *S. poulsonii* genome (Supplementary Data). This contrasts with *Wolbachia*, whose genomes encode more than 20 ankyrin-repeat proteins[21,22]. Notably, the *Wolbachia* strain *w*Mel has a prophage-associated gene *WD0633*, that contains ankyrin repeats and the OTU domain[23]. Nevertheless, its transgenic expression in *D. melanogaster* produced no obvious phenotypes in a previous report[24]. Recent studies discovered *Wolbachia* genes that cause cytoplasmic incompatibility, a reproductive manipulation whereby symbiont-free females are unable to reproduce when mated with infected males[25–27]. Some of the genes that cause cytoplasmic incompatibility along with Spaid contain deubiquitinase domains, raising the possibility that host ubiquitination pathways are a common target in distinct strategies. Future research will focus on the identification of host cellular targets of Spaid, thereby deciphering the principles of its sex-specific activity. A thorough understanding of the reproductive manipulations induced by symbionts would not only provide novel insights into fundamental aspects of development, sex determination, and their evolution in insects, but could also provide clues to control insect populations.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0086-2.

Received: 9 November 2017; Accepted: 21 March 2018;
Published online 2 May 2018.

1. Hurst, G. D. D. & Frost, C. L. Reproductive parasitism: maternally inherited symbionts in a biparental world. *Cold Spring Harb. Perspect. Biol.* **7**, a017699 (2015).
2. Williamson, D. L. & Poulson, D. F. in *The Mycoplasmas, Volume III: Plant and Insect Mycoplasmas* (eds Whitcomb R.F. & Tully J.G.) Ch. 6, 175–208 (Academic, New York, 1979).
3. Williamson, D. L. et al. *Spiroplasma poulsonii* sp. nov., a new species associated with male-lethality in *Drosophila willistoni*, a neotropical species of fruit fly. *Int. J. Syst. Bacteriol.* **49**, 611–618 (1999).
4. Haselkorn, T. S. The *Spiroplasma* heritable bacterial endosymbiont of *Drosophila*. *Fly (Austin)* **4**, 80–87 (2010).
5. Tsuchiyama-Omura, S., Sakaguchi, B., Koga, K. & Poulson, D. F. Morphological features of embryogenesis in *Drosophila melanogaster* infected with a male-killing. *Spiroplasma. Zool. Sci.* **5**, 375–383 (1988).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

6. Veneti, Z., Bentley, J. K., Koana, T., Braig, H. R. & Hurst, G. D. D. A functional dosage compensation complex required for male killing in *Drosophila*. *Science* **307**, 1461–1463 (2005).
7. Bentley, J. K., Veneti, Z., Heraty, J. & Hurst, G. D. D. The pathology of embryo death caused by the male-killing *Spiroplasma* bacterium in *Drosophila nebulosa*. *BMC Biol.* **5**, 9 (2007).
8. Martin, J., Chong, T. & Ferree, P. M. Male killing *Spiroplasma* preferentially disrupts neural development in the *Drosophila melanogaster* embryo. *PLoS One* **8**, e79368 (2013).
9. Harumoto, T., Anbutsu, H. & Fukatsu, T. Male-killing *Spiroplasma* induces sex-specific cell death via host apoptotic pathway. *PLoS Pathog.* **10**, e1003956 (2014).
10. Cheng, B., Kuppanda, N., Aldrich, J. C., Akbari, O. S. & Ferree, P. M. Male-killing *Spiroplasma* alters behavior of the dosage compensation complex during *Drosophila melanogaster* embryogenesis. *Curr. Biol.* **26**, 1339–1345 (2016).
11. Harumoto, T., Anbutsu, H., Lemaitre, B. & Fukatsu, T. Male-killing symbiont damages host's dosage-compensated sex chromosome to induce embryonic apoptosis. *Nat. Commun.* **7**, 12781 (2016).
12. Oishi, K. Spirochaete-mediated abnormal sex-ratio (SR) condition in *Drosophila*: a second virus associated with spirochaetes and its use in the study of the SR condition. *Genet. Res.* **18**, 45–56 (1971).
13. Makarova, K. S., Aravind, L. & Koonin, E. V. A novel superfamily of predicted cysteine proteases from eukaryotes, viruses and *Chlamydia pneumoniae*. *Trends Biochem. Sci.* **25**, 50–52 (2000).
14. Al-Khodor, S., Price, C. T., Kalia, A. & Abu Kwaik, Y. Functional diversity of ankyrin repeats in microbial proteins. *Trends Microbiol.* **18**, 132–139 (2010).
15. Masson, F., Calderon Copete, S., Schüpfer, F., Garcia-Arraez, G. & Lemaitre, B. In vitro culture of the insect endosymbiont *Spiroplasma poulsonii* highlights bacterial genes involved in host-symbiont interaction. *MBio* **9**, e00024-18 (2018).
16. Paredes, J. C. et al. Genome sequence of the *Drosophila melanogaster* male-killing *Spiroplasma* strain MSRO endosymbiont. *MBio* **6**, e02437–14 (2015).
17. Campos, A. R., Rosen, D. R., Robinow, S. N. & White, K. Molecular analysis of the locus *elav* in *Drosophila melanogaster*: a gene whose embryonic expression is neural specific. *EMBO J.* **6**, 425–431 (1987).
18. Salz, H. K. & Erickson, J. W. Sex determination in *Drosophila*: The view from the top. *Fly (Austin)* **4**, 60–70 (2010).
19. Lucchesi, J. C. & Kuroda, M. I. Dosage compensation in *Drosophila*. *Cold Spring Harb. Perspect. Biol.* **7**, a019398 (2015).
20. Madigan, J. P., Chotkowski, H. L. & Glaser, R. L. DNA double-strand break-induced phosphorylation of *Drosophila* histone variant H2Av helps prevent radiation-induced apoptosis. *Nucleic Acids Res.* **30**, 3698–3705 (2002).
21. Siozios, S. et al. The diversity and evolution of *Wolbachia* ankyrin repeat domain genes. *PLoS ONE* **8**, e55390 (2013).
22. Jernigan, K. K. & Bordenstein, S. R. Ankyrin domains across the Tree of Life. *PeerJ* **2**, e264 (2014).
23. Wu, M. et al. Phylogenomics of the reproductive parasite *Wolbachia pipientis* *w*Mel: a streamlined genome overrun by mobile genetic elements. *PLoS Biol.* **2**, e69 (2004).
24. Yamada, R., Iturbe-Ormaetxe, I., Brownlie, J. C. & O'Neill, S. L. Functional test of the influence of *Wolbachia* genes on cytoplasmic incompatibility expression in *Drosophila melanogaster*. *Insect Mol. Biol.* **20**, 75–85 (2011).
25. Beckmann, J. F. & Fallon, A. M. Detection of the *Wolbachia* protein WPIP0282 in mosquito spermathecae: implications for cytoplasmic incompatibility. *Insect Biochem. Mol. Biol.* **43**, 867–878 (2013).
26. Beckmann, J. F., Ronau, J. A. & Hochstrasser, M. A. *Wolbachia* deubiquitylating enzyme induces cytoplasmic incompatibility. *Nat. Microbiol.* **2**, 17007 (2017).
27. LePage, D. P. et al. Prophage WO genes recapitulate and enhance *Wolbachia*-induced cytoplasmic incompatibility. *Nature* **543**, 243–247 (2017).

Acknowledgements We thank the Bloomington Stock Center in the USA and the Department of *Drosophila* Genomics and Genetic Resources at Kyoto Institute of Technology in Japan (DGGR) for fly stocks; the Developmental Studies Hybridoma Bank at the University of Iowa for monoclonal antibodies; the *Drosophila* Genomics Resource Center in the USA for DNA vectors; J. Jaenike, T. Murata, and M. Kuroda for providing fly strains; T. Fukatsu for the *S. poulsonii*-infected fly stocks that were originally established in his laboratory by T.H; J. Lucchesi for providing antibodies; the BioImaging & Optics Platform (BIOP) in EPFL for confocal microscopy; K. Harshman and the Lausanne Genomic Technologies Facility (GTF) at the University of Lausanne (UNIL) for whole-genome sequencing and draft genome assemblies; and C. Vorburger and S. Perlman for comments and suggestions on the manuscript; and the members of the laboratory for discussion and support. This work was supported by the European Research Council (ERC) Advanced Grant 339970 and the Swiss National Science Foundation (SNSF) Sinergia Grant CRSII3_154396. The purchase of the PacBio RS II system in UNIL was supported in part by the Loterie Romande through the Fondation pour la Recherche en Médecine Génétique.

Reviewer information *Nature* thanks P. Ferree and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.H. conceived of and designed the study, performed the experiments, analysed the data and wrote and edited the manuscript. B.L. conceived and administered the study and wrote and edited the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0086-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0086-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.H. or B.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Fly stocks and genetics. Laboratory stocks of *D. melanogaster* were maintained at 25 °C with standard cornmeal medium. The fly stocks used in this study were cultured in tetracycline-containing medium (0.7–0.8 mg ml^{-1}) for one generation to eliminate possible contamination with *Wolbachia* and *Spiroplasma*. After treatment, about ten females and/or males were checked by diagnostic PCR with specific primers for *Wolbachia* (wsp_81F/691R)[28] and *Spiroplasma* (16SA1F/TKSSsp)[29,30] (Supplementary Table 1). Treated flies were maintained in the normal medium for at least two generations before use in experiments. The following lines were obtained from the Bloomington *Drosophila* Stock Center at Indiana University (BDSC) and the Department of *Drosophila* Genomics and Genetic Resources at Kyoto Institute of Technology (DGGR): *actin5C–GAL4* (*actin–GAL4*; BDSC 4414), *nanos–GAL4::VP16* (*nanos–GAL4*; BDSC 4937), *tubulin–GAL80^{ts}* (BDSC 7108), *armadillo–GAL4* (BDSC 1560), *UASp–eGFP* (DGGR 116071), and *CyO, ActGFP* (the green balancer; DGGR 107783). The *nanos–GAL4* flies, which were found to be *Wolbachia* and *Spiroplasma*-free by PCR, were not treated with antibiotics, because they became sick after treatment. Oregon-R (used as a wild-type line) and *msl1^{L60}/CyO*; *H83M2*[31] flies were generously provided by T. Murata and M. Kuroda, respectively.

To express the Spaid-encoding gene, we used the GAL4/UAS system[32] (see below). For the zygotic expression of Spaid, *actin–GAL4/CyO* (Fig. 1c) or *armadillo–GAL4* (homozygous; Extended Data Fig. 6) flies were crossed to homozygous *UAS* transgenic flies. To determine the lethal phase during larval stages, we analysed the number of teeth present in mouth hooks of killed larvae[33]. For maternal expression[34], *UAS* females were crossed to *nanos–GAL4* males, and the resultant female progeny were mated with Oregon-R males (Figs. 2, 3 and Extended Data Fig. 4 and 5e, f). For the ectopic formation of the MSL complex in females, *msl1^{L60}/CyO, ActGFP*; *H83M2* flies were used instead of Oregon-R. The *H83M2* transgene expresses the *msl-2* coding sequence (lacking Sxl-binding sites and resistant to translational repression) under the control of a heat shock-inducible promoter[31], but we used its leaky basal expression. The resultant embryos were distinguished by GFP staining and only GFP$^+$ embryos (wild-type for *msl1*) were analysed (Extended Data Fig. 5a–d). For the expression of Spaid in larval salivary glands (Fig. 4), we used a recombined *actin–GAL4, tubulin–GAL80^{ts}/CyO* line to avoid male lethality and let them grow until the third instar larval stage. Crosses maintained at 18 °C for 7–8 days were shifted to 29 °C and kept for 1 day before dissection. Only GFP$^+$ wandering third instar larvae were dissected.

Construction of transgenic fly lines. For the gene synthesis of the Spaid coding sequence by the GeneArt service (Thermo Fisher Scientific), TGA stop codons, which encode tryptophan in *Spiroplasma*, were mutated into TGG. Codon usage was optimized for the expression in *D. melanogaster*. Consequently, GC content was modified from 23.3% to 50.4%. We divided the Spaid coding sequence into two parts and synthesized them separately: one was a 2,367-bp fragment that corresponds to Spaid ΔC and the other was an 838-bp fragment encoding the remaining 3′ portion of full-length Spaid. To obtain the full-length Spaid coding sequence (3,195 bp, without a stop codon), the two fragments were fused using PCR with a 30-bp overlap and cloned into the pENTR vector using the pENTR/D-TOPO cloning kit (Thermo Fisher Scientific). To generate Spaid deletion constructs, we amplified two PCR fragments (nucleotide positions 1–276 and 787–3195 for ΔANK; nucleotide positions 1–789 and 1396–3195 for ΔOTU) from the synthetic Spaid coding sequence, fused them using PCR with a 24-bp overlap and cloned the product into the pENTR vector. We used PrimeSTAR HS DNA Polymerase (Takara Bio) for all PCR reactions. The Gateway cassette containing the Spaid coding fragments was transferred into the pPWG destination vector (The *Drosophila* Genomics Resource Center 1078; The *Drosophila* Gateway Vector Collection) by the LR clonase II enzyme mix kit (Thermo Fisher Scientific) to construct pUASp-Spaid-eGFP, pUASp-Spaid.ΔANK-eGFP, and pUASp-Spaid.ΔOTU-eGFP plasmids. Transgenic fly lines were generated by the standard microinjection method for P-element transformation (BestGene).

Identification and characterization of the partial male-killing *S. poulsonii* strain. Male-killing *S. poulsonii* can be easily transferred to other fly stocks by haemolymph injection[2]. *S. poulsonii* containing haemolymph was collected from a naturally infected *D. melanogaster* line Ug-SR provided by J. Jaenike[35], and transferred to wild-type Oregon-R flies that were used as the source of haemolymph for subsequent artificial infection. We established two artificially infected *D. melanogaster* stocks, *Df(3L)H99* (DGGR 106395) and *Sxl-eGFP* (BDSC 24105) in 2012 (Extended Data Fig. 1a). These infected stocks had shown perfect male killing (no emergence of adult males) for a period of at least 20 generations after the establishment of stable infection[9]. Afterwards, however, the infected *Sxl-eGFP* line started to produce male escapers, while the *Df(3L)H99* line showed complete male killing. We collected haemolymph from these fly stocks and injected them into Oregon-R flies to exclude the effect of host genetic background. We confirmed that only the *S. poulsonii* strain collected from the *Sxl-eGFP* line showed partial male killing (Extended Data Fig. 1b, c). The male-killing *S. poulsonii* strain in *D. melanogaster* is conventionally designated as MSRO (*melanogaster* sex ratio organism). Therefore, we call these *S. poulsonii* strains MSRO-H99 (the complete male killer as a control) and MSRO-SE (the partial male killer) after the genotypes of the host fly lines from which they were derived (*Df(3L)H99* and *Sxl-eGFP*, respectively). We also refer to the original male-killing strain from the Ug-SR line as MSRO-Ug according to the previous study[35].

We suspect that the partial male-killing strain might have resulted from genome rearrangements as reported in other *Spiroplasma* species maintained in the laboratory[36]. Repeat-rich sequences like viral fragments are often associated with genomic instability of *Spiroplasma*[37], and accordingly, the *spaid* gene of the partial male-killing strain seemed to be truncated by rearrangements between genes containing repetitive sequences (for example, *integrase* and *p58*; Extended Data Fig. 3b). To keep the partial male-killing strain, we used aged females (7–11 days) harbouring high bacterial titres to prevent the loss of the symbiont during vertical transmission. Because the titre of the partial male-killing strain appears to be higher than the original strain in some aged females, this may have led to the propagation of mutant bacteria and promoted their transmission (Extended Data Fig. 1d).

Analysis of *S. poulsonii* titres by quantitative PCR. Oregon-R virgin females infected with MSRO-Ug, MSRO-H99, and MSRO-SE were collected and aged for 0 (virgin), 7 or 14 days and individually transferred to a 1.5-ml tube and stored at −80 °C. They were then homogenized with plastic pestles by hand, and genomic DNA was purified using the DNeasy Blood & Tissue Kit (Qiagen). A portion of the *S. poulsonii dnaA* gene and the *D. melanogaster RpS17* gene were amplified with specific primer sets (dnaA_109F/246R[38] and rps17_615F/695R[39], respectively) (Supplementary Table 1) under the following PCR conditions: 95 °C for 5 min and then 45 cycles of 95 °C for 10 s, 55 °C for 20 s and 72 °C for 20 s. The 10-μl PCR mixture contained 2 μl of genomic DNA, 1× LightCycler 480 SYBR Green I master mix (Roche) and 0.5 μM primers. Quantitative PCR was performed by LightCycler 480 II (Roche). Cp values were obtained by the second derivative maximum method and the values of technical duplicates were averaged. We excluded samples if their duplicates had a difference in Cp values >0.5 cycles or if their Cp values were more than 30 cycles (2 out of 63 samples in total). To calculate the relative copy number of *S. poulsonii dnaA* gene against the host *RpS17* gene, we followed the Pfaffl method[40]. To estimate the PCR efficiency of each primer set, we performed quantitative PCR by using six tenfold serial dilutions of genomic DNA purified from ten female adult flies infected with MSRO-Ug, aged for 4–6 days after eclosion. PCR efficiency values were 82% (dnaA_109F/246R) and 98% (rps17_615F/695R), respectively.

Whole-genome sequencing of male-killing *S. poulsonii*. To collect a sufficient amount of DNA for whole-genome sequencing of *S. poulsonii* strains MSRO-H99 and MSRO-SE, we recovered fly haemolymph by a previously developed centrifugal separation method[41] with minor modifications. We first prepared a 0.5-ml polypropylene tube (Sarstedt, 72.704) whose bottom has a slit made by a sharp blade (hereafter called cartridge). Under CO_2 anaesthesia, we pricked the preepisternal area of the thorax of infected female adults and put them into the cartridge. The cartridge containing 30–40 flies was inserted into a 1.5-ml tube and centrifuged at 5,000 r.p.m. (2,300g) for 5 min at 4 °C. We usually made between three and ten cartridges at one time and they were kept on ice until centrifugation. After removing cartridges, we checked precipitates that contained host haemolymph and *S. poulsonii* under a stereomicroscope, and confirmed that there were no large debris (embryos and carcasses etc.). The tubes were washed with 400 μl PBS to resuspend and merge all precipitates. The merged suspension was filtered with a 0.65-μm pore size membrane filter (Ultrafree-MC; Merck, UFC30DV00) by centrifugation at 12,000g for 3 min at 4 °C to eliminate large microorganisms including yeasts. The flow-through was centrifuged at maximum speed for another 10 min and the supernatant was removed. Remaining bacterial pellets were stored at −80 °C. After suspending the bacterial pellet in 180 μl buffer ATL (Qiagen), 20 μl proteinase K solution (Qiagen) was added and pulse-vortexed for 10 s, and the mixture was then incubated for 1 h at 56 °C. Then, 4 μl RNase A (100 mg ml^{-1}; Qiagen) was added and treated for 5 min at room temperature. Genomic DNA was purified by the conventional phenol–chloroform extraction method (see PacBio Shared Protocol online, https://www.pacb.com/wp-content/uploads/2015/09/SharedProtocol-Extracting-DNA-usinig-Phenol-Chloroform.pdf). During the extraction, samples were mixed by pulse-vortexing within 20 s to prevent shearing of high molecular weight DNA. We collected haemolymph from 1,843 (MSRO-H99) and 1,378 (MSRO-SE) adult females to recover 18.9 μg and 7.56 μg genomic DNA, respectively (quantified by NanoDrop 1000; Thermo Fisher Scientific).

The genomic DNA was purified with Agencourt AMPure XP beads (Beckman Coulter) and was sheared in a Covaris g-TUBE (Covaris) to obtain 20-kb fragments. After shearing, the size distribution of DNA was checked by the Fragment Analyzer (Advanced Analytical Technologies). Sheared DNA (4 μg, MSRO-H99; 5 μg MSRO-SE) was used to prepare a SMRTbell library with the PacBio SMRTbell

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Template Prep Kit 1.0 (Pacific Biosciences) according to the manufacturer's recommendations. The resulting library was size selected on the BluePippin system (Sage Science) for molecules larger than 16 kb (MSRO-H99) or 18 kb (MSRO-SE). The recovered library bound to MagBeads was sequenced on a single SMRT Cell with P6/C4 chemistry per genome by the PacBio RS II system (Pacific Biosciences) at 240-min movie length. Genome assembly was performed with the HGAP (Hierarchical Genome Assembly Process) software (Pacific Biosciences) version 3 and version 2 for the MSRO-H99 and MSRO-SE genomes, respectively. Library preparation, whole-genome sequencing and genome assembly were performed in the Lausanne Genomic Technologies Facility (GTF) at the University of Lausanne (UNIL).

Genomic data analyses and protein domain searches. Genomic data analysis was performed on the Bio-Linux 8 platform (NERC Environmental Omics)[42]. The genome sequences of MSRO-H99 and MSRO-SE assembled into five and two contigs, respectively (Extended Data Fig. 2a), and were aligned with the previously published MSRO genome[16] and ordered using Mauve v.2.4.0[43,44] (Extended Data Fig. 2b). For the MSRO-H99 genome, three major contigs (1–3) were assigned to the main chromosome, while the remaining two minor contigs (4 and 5) were assigned to extra chromosomes. Two contigs of the MSRO-SE genome were assigned to the main chromosome (contig 1; circularized) and the extra chromosome (contig 2). These extra chromosomes could be plasmids because they contain several proteins, such as Soj (the chromosome partitioning protein) and ARPs (adhesion-related proteins; P58, P12, P54, P123 and P18), which are located on plasmids in other *Spiroplasma* species[45].

Whole-genome annotation was conducted by using Prokka v.1.11[46]. We created a custom annotation database by combining published genomic sequences of several *Spiroplasma* species (*S. citri* CII3-3X (AM285301)[47], *S. melliferum* KC3 (AGBZ02000000)[48], *S. melliferum* IPMB4A (AMGI01000000)[49], *S. chrysopicola* DF-1 (CP005077)[37], *S. poulsonii* MSRO (NZ_JTLV00000000)[16], *S. kunkelii* CR2-3X (CP010899)[50]; the numbers in parentheses represent GenBank accession numbers).

To identify conserved protein domains including ankyrin repeats, annotated protein sequences were analysed using the NCBI Conserved Domain Database (NCBI CDD)[51,52] with default parameters (Supplementary Data). Besides *spaid* (SMH99_26490; 3,198 bp), other genes containing the OTU domain (SMH99_25890; 597 bp) and ankyrin repeats (SMH99_25900; 204 bp) were predicted on an identical contig (contig 4; 87,892 bp) in MSRO-H99. It is likely that they are derived from the misassembly or partial assembly of Spaid-encoding reads, because they are located close to the end of the contig (from 4,133 to 5,048 bp) with much lower coverage (below 10×) compared to the *spaid* locus (from 51,206 to 54,403 bp) with high coverage (over 330×). These genes were not found in MSRO-SE. Further domains were predicted using the InterPro protein sequence analysis and classification database[53]. The protein domain structure of Spaid was drawn using the Illustrator for Biological Sequences (IBS) software v 1.0.2[54] and modified manually.

Homology searches. For homology searches, we used the protein BLAST (blastp) program on the NCBI BLAST website with a non-redundant protein sequence database (nr). After searching the entire database without specifying the organism, we also searched bacteria, viruses, and *Drosophila* databases separately; however, we found no protein sequences that aligned to the entire Spaid protein sequence. The top hit having the highest score was a hypothetical protein of *Spiroplasma kunkelii* (WP_053391598), which covers the portion of the N- and C-terminal sequences of Spaid (amino acid positions 1–49 (score, 45.4; *E*-value, 0.71; identities, 48.98%; positives, 69.39%; gaps, 14.29%) and 704–1,065 (score, 385; *E*-value, 2.63×10^{-118}; identities, 58.03%; positives, 69.63%; gaps, 14.07%), respectively).

Sanger sequencing of the *spaid* locus. Ten adult Oregon-R females that were uninfected (as a negative control) or infected with MSRO-Ug, MSRO-H99, and MSRO-SE were collected and genomic DNA was purified using the method described above for quantitative PCR. The 3′ portion of the *spaid* gene was amplified with forward and reverse primer sets (*spaid*_1568F with *spaid*.L_ +136R for MSRO-Ug and MSRO-H99 and *spaid*.S_ +362R for MSRO-SE, respectively) (Supplementary Table 1) under the following PCR conditions: 95 °C for 2 min then 30 cycles of 95 °C for 30 s, 50 °C for 30 s and 60 °C for 3 min. Owing to the low GC content of the *spaid* gene (23.3%), extension at 60 °C instead of 72 °C was essential for successful PCR amplification[55]. The 50-μl PCR mixture contained 2 μl of genomic DNA, 1.25 U of GoTaq G2 DNA polymerase (Promega), 1× green reaction buffer (Promega), 0.5 μM primers, and 0.2 mM dNTP mixture. The PCR fragment was purified from agarose gel using the Wizard SV Gel and PCR Clean-Up System (Promega) and read by direct sequencing. To design primers, we referred to the genome sequences assembled from the PacBio data. No distinct PCR amplification was detected in uninfected Oregon-R female flies, confirming that the *spaid* gene is not encoded by the *D. melanogaster* genome but by the *S. poulsonii* genome.

Staining and imaging. To collect embryos, we put female and male flies in culture vials and waited for two days. These adults were transferred to egg-collection cages with grape-juice agar plates and fed with yeast paste. Developmental staging of embryos was followed refs. 56,57. Embryos were collected from grape-juice agar plates and dechorionated in 2.8% sodium hypochlorite solution, they were subsequently fixed in 1:1 mixture of heptane and 4% paraformaldehyde (EM Grade; Electron Microscopy Sciences, 15710) diluted in PBS for 20 min at room temperature, and devitellinized by vigorously shaking in heptane/MeOH. The embryos were washed in MeOH three times and rehydrated through an EtOH series (95%, 70%, 50% and 35%), and then washed in PBT (PBS containing 0.1% Triton X-100). After treated with a blocking buffer (PBT containing 2% bovine serum albumin (BSA, Fraction V; MP Biomedicals, 02160069)) for 60 min, the embryos were incubated with primary antibodies at 4 °C overnight, washed three times in PBT and incubated with secondary antibodies at room temperature for 90 min. Antibodies were diluted in the blocking buffer. Anti-Sxl and anti-MSL1 antibodies were used for sexing embryos (see below).

The following primary antibodies were used: mouse anti-Sxl (1:20; Developmental Studies Hybridoma Bank (DSHB), M18)[58], mouse and rabbit anti-MSL1 (1:200 and 1:500; kindly provided by J. Lucchesi), rabbit anti-histone H2AvD pS137 (pH2Av; 1:1,000; Rockland Immunochemicals, 600-401-914), chicken anti-GFP (1:500; Aves Labs, GFP-1020), rat anti-Elav (1:20; DSHB, 7E8A10)[59]. Secondary antibodies (1:1,000; Alexa Fluor 488/555/647 conjugate) were purchased from Molecular Probes (Thermo Fisher Scientific). DNA staining was carried out with DAPI (1 μg ml^{-1}; Sigma, D9542) for 10 min at room temperature after secondary antibody staining. TUNEL staining was performed by using the In situ Cell Death Detection Kit, TMR red (Roche), and the embryos stained with primary antibodies were incubated in 50 μl TUNEL reaction mixture with secondary antibodies at 4 °C overnight. Stained embryos were washed three times in PBT, mounted in FluorSave Reagent (Calbiochem) and observed under a confocal microscope (Zeiss LSM 700).

To detect *S. poulsonii* in host haemolymph (Fig. 1a), the abdomen of an adult female was dissected in 5 μl PBS with tweezers on a microscope slide. One microlitre of SYTO 9 nucleic acid stain solution (0.02 mM; Thermo Fisher Scientific) was added and observed under a widefield microscope equipped with a CCD camera (Zeiss Axio Imager Z1/AxioCam MRm).

To monitor the subcellular distribution of the GFP fusion proteins of Spaid, we dissected out salivary glands from wandering third instar larvae and fixed them in 4% paraformaldehyde diluted in PBS for 15 min at room temperature. Blocking and staining were performed as described above. For GFP, we detected raw fluorescent signals without antibody staining.

Image analysis and processing. Confocal *z*-sections were max projected using the Fiji software package[60] with a custom macro. Image analysis was performed using custom R scripts and the EBImage package[61]. For the quantification of TUNEL signals of whole embryos (Fig. 2c and Extended Data Fig. 5d) (acquired using a 20×/0.8 objective with 0.6× zoom scan; frame size: 512 × 512; 1.5 μm, two times optimal intervals), maximum projection images of DAPI and TUNEL staining were binarized and the former was used to make an embryonic mask image. TUNEL signals inside a corresponding mask image were measured by image integration. This value was divided by mask image area to normalize the embryonic size.

For the quantification of focal pH2Av signals (Fig. 3c), embryos from stages 8–10 were triply stained (pH2Av, MSL1, and DAPI) and two images were acquired per embryo (using a 63×/1.4 oil immersion objective; frame size, 1,024 × 1,024; 0.3 μm optimal intervals). We compiled 20 serial *z*-sections from the top (from 25 sections in total) to make projected images of pH2Av and MSL1 signals. These images were smoothed using a Gaussian filter and binarized using the moving average method, respectively. To identify focal pH2Av signals, image objects were extracted from the stacked images by segmentation and labelling (objects that were smaller than 20 pixels were eliminated to remove noise and ambiguous signals). To calculate the enrichment of focal pH2Av signals on the MSL1-labelled chromosome (Fig. 3d), overlaps between focal pH2Av signals and MSL1 signals were obtained by image integration. To quantify the number of MSL1 signals with discrete focal shapes (Fig. 3h), objects over 30 pixels were extracted and counted. The data sets described above were also used to count the number of chromatin bridges (Fig. 3g). The 25 serial *z*-slice images were visually inspected to detect chromatin bridges in the Fiji software.

The brightness and contrast of the presented images in the manuscript were adjusted using the level tool in the GIMP 2.8 software package. The adjustment was performed uniformly on the entire images and only black/white input levels were modified.

Statistics and reproducibility. No statistical methods were applied to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. We used the R v.3.4.1 software package (the R Foundation) for all statistical analyses. Multiple comparisons in Figs. 1c, 2c, 3c and Extended Data Fig. 1d, 5d were performed using the Steel–Dwass test of the pSDCFlig function in the NSM3 R package[62]. We performed the χ^2 test in Fig. 3d and the Mann–Whitney U test (two-tailed)

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

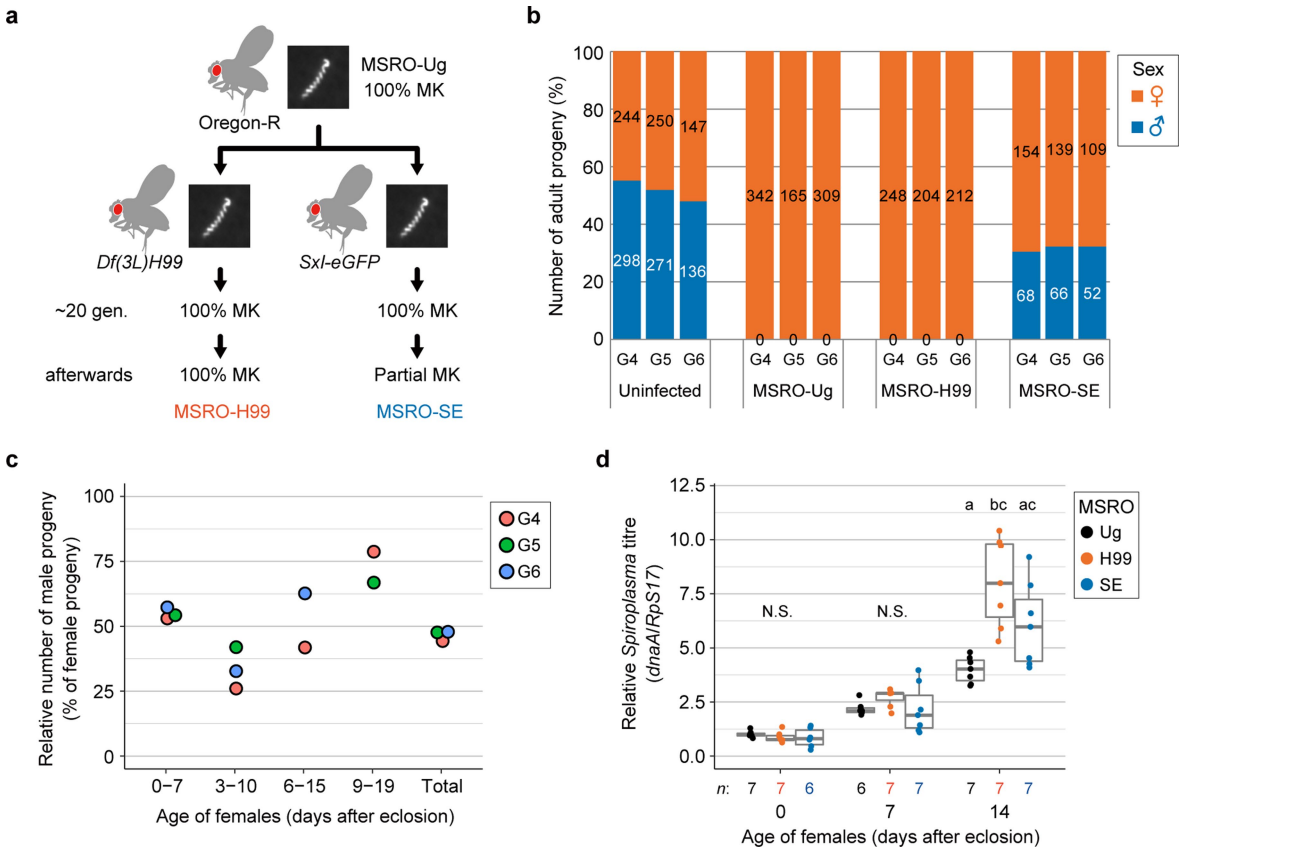
in Fig. 3h and Extended Data Fig. 6. *P*-values less than 0.05 were considered as significant. Exact *P*-values are listed in Supplementary Table 2. For the zygotic expression of *spaid* and its deletion constructs (Fig. 1c and Extended Data Fig. 6), two independently established *UAS* transgenic lines were used to verify the reproducibility of the results. Other experiments were independently repeated at least two times, except for the qPCR analysis in Extended Data Fig. 1d (one experiment with 6–7 biological replicates).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Whole genomic sequence data have been deposited at GenBank under the BioProject number PRJNA416288. Sequence data for synthetic gene fragments have been deposited at GenBank under the accession numbers MG837001 and MG837002. Source Data for Figs. 1, 2, 3 and Extended Data Fig. 1, 5, 6 are available in the online version of the paper. All relevant data supporting the findings of this study are included within the article and its Supplementary Information Files or are available from the corresponding authors upon reasonable request.

28. Zhou, W., Rousset, F. & O'Neil, S. Phylogeny and PCR-based classification of *Wolbachia* strains using *wsp* gene sequences. *Proc. Biol. Sci.* **265**, 509–515 (1998).
29. Fukatsu, T. & Nikoh, N. Two intracellular symbiotic bacteria from the mulberry psyllid *Anomoneura mori* (Insecta, Homoptera). *Appl. Environ. Microbiol.* **64**, 3599–3606 (1998).
30. Fukatsu, T. & Nikoh, N. Endosymbiotic microbiota of the bamboo pseudococcid *Antonina crawii* (Insecta, Homoptera). *Appl. Environ. Microbiol.* **66**, 643–650 (2000).
31. Kelley, R. L. et al. Expression of msl-2 causes assembly of dosage compensation regulators on the X chromosomes and female lethality in *Drosophila*. *Cell* **81**, 867–877 (1995).
32. Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
33. Ashburner, M., Golic, K. G. & Hawley, R. S. *Drosophila: a Laboratory Handbook. 2nd Edition*, (Cold Spring Harbor, New York, 2005).
34. Rørth, P. Gal4 in the *Drosophila* female germline. *Mech. Dev.* **78**, 113–118 (1998).
35. Pool, J. E., Wong, A. & Aquadro, C. F. Finding of male-killing *Spiroplasma* infecting *Drosophila melanogaster* in Africa implies transatlantic migration of this endosymbiont. *Heredity (Edinb)* **97**, 27–32 (2006).
36. Ye, F., Melcher, U., Rascoe, J. E. & Fletcher, J. Extensive chromosome aberrations in *Spiroplasma citri* Strain BR3. *Biochem. Genet.* **34**, 269–286 (1996).
37. Ku, C., Lo, W.-S., Chen, L.-L. & Kuo, C.-H. Complete genomes of two dipteran-associated spiroplasmas provided insights into the origin, dynamics, and impacts of viral invasion in *Spiroplasma*. *Genome Biol. Evol.* **5**, 1151–1164 (2013).
38. Anbutsu, H. & Fukatsu, T. Population dynamics of male-killing and non-male-killing spiroplasmas in *Drosophila melanogaster*. *Appl. Environ. Microbiol.* **69**, 1428–1434 (2003).
39. Osborne, S. E., Leong, Y. S., O'Neill, S. L. & Johnson, K. N. Variation in antiviral protection mediated by different *Wolbachia* strains in *Drosophila simulans*. *PLoS Pathog.* **5**, e1000656 (2009).
40. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Res.* **29**, e45 (2001).
41. Musselman, L. P. *Drosophila* hemolymph collection procedure. *YouTube* https://www.youtube.com/watch?v=im78OIBKIPA (2013).
42. Field, D. et al. Open software for biologists: from famine to feast. *Nat. Biotechnol.* **24**, 801–803 (2006).
43. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
44. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
45. Saillard, C. et al. The abundant extrachromosomal DNA content of the *Spiroplasma citri* GII3-3X genome. *BMC Genomics* **9**, 195 (2008).
46. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
47. Carle, P. et al. Partial chromosome sequence of *Spiroplasma citri* reveals extensive viral invasion and important gene decay. *Appl. Environ. Microbiol.* **76**, 3420–3426 (2010).
48. Alexeev, D. et al. Application of *Spiroplasma melliferum* proteogenomic profiling for the discovery of virulence factors and pathogenicity mechanisms in host-associated spiroplasmas. *J. Proteome Res.* **11**, 224–236 (2012).
49. Lo, W.-S., Chen, L.-L., Chung, W.-C., Gasparich, G. E. & Kuo, C.-H. Comparative genome analysis of *Spiroplasma melliferum* IPMB4A, a honeybee-associated bacterium. *BMC Genomics* **14**, 22 (2013).
50. Davis, R. E. et al. Complete genome sequence of *Spiroplasma kunkelii* strain CR2-3X, causal agent of corn stunt disease in *Zea mays* L. *Genome Announc.* **3**, e01216–15 (2015).
51. Marchler-Bauer, A. et al. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229 (2011).
52. Marchler-Bauer, A. et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222–D226 (2015).
53. Finn, R. D. et al. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* **45**, D190–D199 (2017).
54. Liu, W. et al. IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* **31**, 3359–3361 (2015).
55. Su, X. Z., Wu, Y., Sifri, C. D. & Wellems, T. E. Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA. *Nucleic Acids Res.* **24**, 1574–1575 (1996).
56. Hartenstein, V. *Atlas of Drosophila Development*. (Cold Spring Harbor, New York, 1993).
57. Campos-Ortega, J. A. & Hartenstein, V. *The Embryonic Development of Drosophila melanogaster*. (Springer, Berlin, 1997).
58. Bopp, D., Bell, L. R., Cline, T. W. & Schedl, P. Developmental distribution of female-specific Sex-lethal proteins in *Drosophila melanogaster*. *Genes Dev.* **5**, 403–415 (1991).
59. O'Neill, E. M., Rebay, I., Tjian, R. & Rubin, G. M. The activities of two Ets-related transcription factors required for *Drosophila* eye development are modulated by the Ras/MAPK pathway. *Cell* **78**, 137–147 (1994).
60. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
61. Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBImage—an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–981 (2010).
62. Hollander, M., Wolfe, A. D. & Chicken, E. *Nonparametric Statistical Methods, 3rd Edition*, (John Wiley & Sons, New Jersey, 2013).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



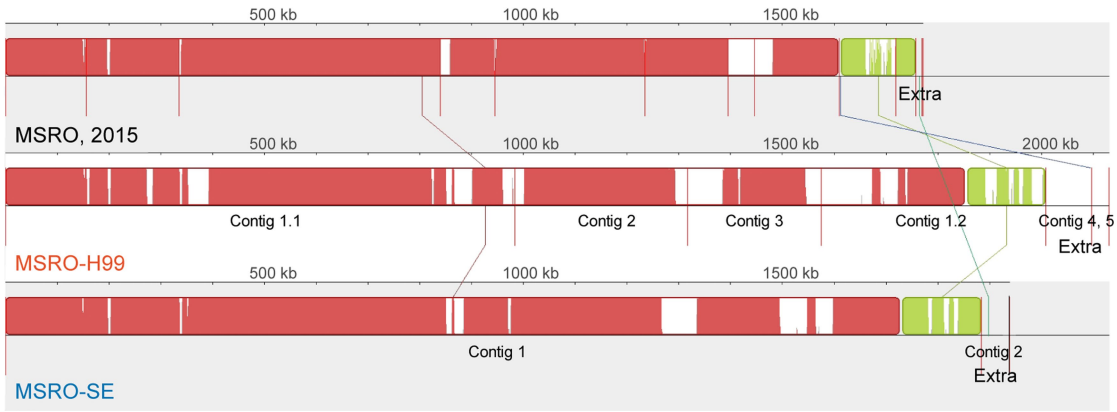
Extended Data Fig. 1 | Identification and characterization of the partial male-killing *S. poulsonii* strain. a, An illustration showing the origin of the *S. poulsonii* strains analysed in this study. Pictures show male-killing (MK) *S. poulsonii* of *D. melanogaster*. MSRO-Ug is the original male-killing strain maintained in the Oregon-R wild-type fly. Fly stocks (*Df(3L)H99* and *Sxl-eGFP*) artificially infected with this original strain showed complete male killing (100% MK) for the first 20 generations. Afterwards, one strain (MSRO-SE) started to show the partial male-killing phenotype, whereas the other (MSRO-H99) kept the ability to induce complete male killing. See Methods section 'Identification and characterization of the partial male-killing *S. poulsonii* strain' for more detail. **b**, **c**, Sex-ratio analysis of the adult progeny obtained from Oregon-R flies infected with MSRO-Ug, MSRO-H99, and MSRO-SE. We repeated experiments three times on the fourth, fifth and sixth generations (G4–6) after the establishment of infection. In **c**, the relative number of male offspring (percentage of females) obtained from Oregon-R female flies infected with MSRO-SE are plotted. Data points were excluded if the total count of flies was below 10. **d**, Relative titre of *S. poulsonii* within individual female flies. Adult females infected with three MSRO strains were kept for 0, 7 and 14 days after eclosion and analysed by qPCR. Data were normalized with respect to females from day 0 that were infected with MSRO-Ug. Different letters indicate statistically significant different groups ($P < 0.01$; N.S., not significant, $P > 0.05$; Steel–Dwass test; see Supplementary Table 2). Please note that the titres of the three strains are comparable and even the higher titre in old females (see 14 days in **d**) fails to induce complete male killing in MSRO-SE (**c**). Box and dot plots are as in Fig. 1c and sample sizes (*n*, number of adult flies) are shown at the bottom.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

a

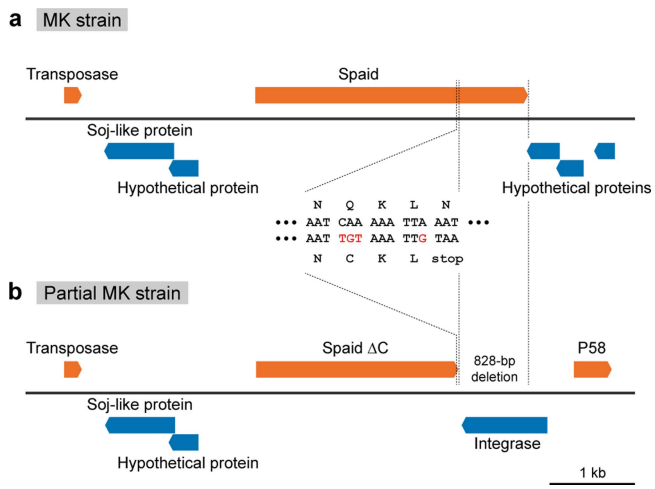
	MSRO, 2015	MSRO-H99	MSRO-SE
Reference	Paredes *et al.*, 2015	This work	This work
Male killing	Complete	Complete	Partial
Contigs	12	5	2
Largest contig (bp)	504,367	1,417,292	1,883,572
Contig 1	-	1,417,292	1,883,572
Contig 2	-	333,653	55,268
Contig 3	-	257,938	-
Contig 4	-	87,892	-
Contig 5	-	34,188	-
Total length (bp)	1,771,859	2,130,963	1,938,840
GC (%)	27	26	26
N50	179,219	1,417,292	1,883,572
N75	155,942	333,653	1,883,572
L50	3	1	1
L75	6	2	1
Total genes	1,976	2,749	2,516
CDS	1,942	2,715	2,482
tRNA	31	31	31
rRNA	3	3	3

b



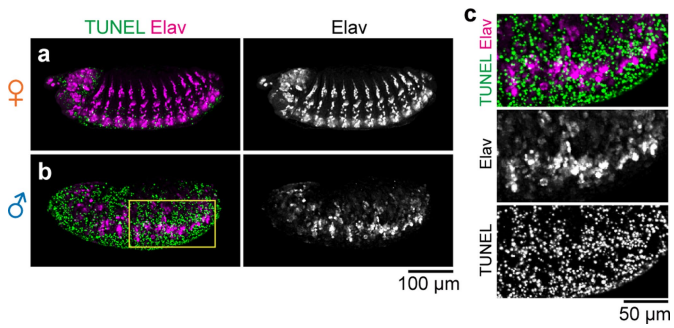
Extended Data Fig. 2 | Whole-genome sequencing studies of the male-killing *S. poulsonii* variants. a, Comparison of the genomic features of three *S. poulsonii* strains. MSRO-H99 and MSRO-SE are newly obtained variants isolated in this study (Extended Data Fig. 1a). As a control, data from the previously reported male-killing *S. poulsonii* genome[16] are also indicated. **b**, Whole-genome alignment of the three *S. poulsonii* strains. To start the alignment from the *dnaA* gene, contig 1 of MSRO-H99 was split into two fragments (contig 1.1 and 1.2). The locations of the contigs corresponding to extra chromosomes (putative plasmids; see Methods section 'Genomic data analyses and protein domain searches') are shown as 'extra'. *spaid* (gene ID from GenBank BioProject PRJNA416288: SMH99_26490) and *spaid* ΔC (gene ID: SMSE_25110) are located on these extra chromosomes in MSRO-H99 (contig 4) and MSRO-SE (contig 2), respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



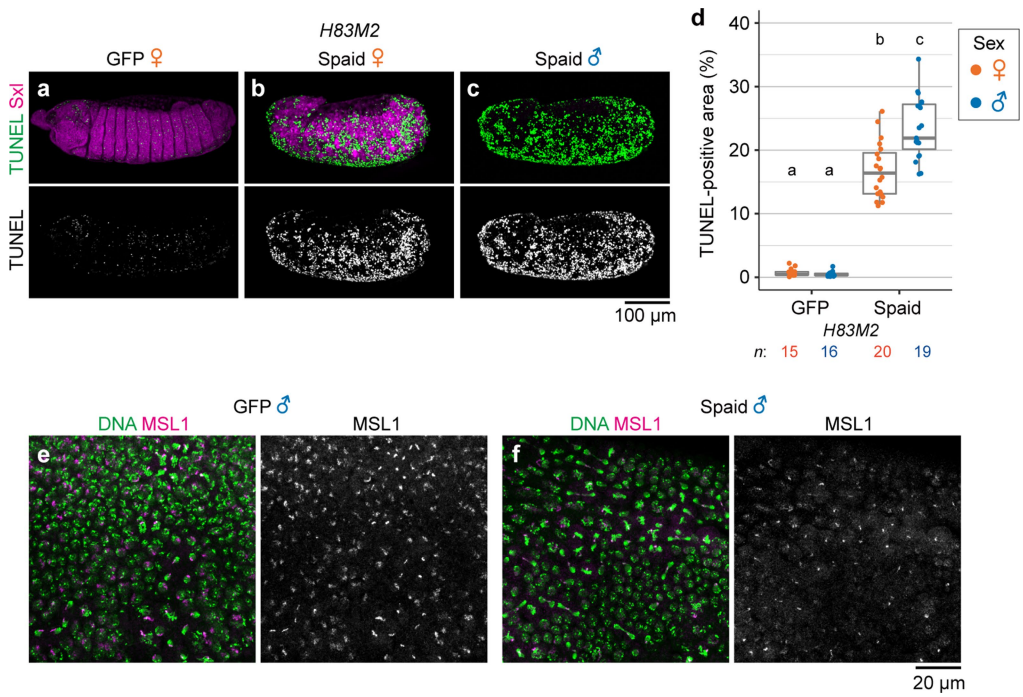
Extended Data Fig. 3 | Genetic alterations of the *spaid* locus in the partial male-killing *S. poulsonii* strain. The genome structures around the *spaid* loci in the male-killing (**a**, MSRO-Ug and MSRO-H99) and the partial male-killing (**b**, MSRO-SE) *S. poulsonii* strains. Genes encoded on opposite strands are shown in different colours (red and blue, respectively). An 828-bp deletion and nucleotide substitutions (coloured in red; corresponding amino acid sequences are presented in one-letter code) in the 3′ region of the *spaid* gene are indicated. These sequence alterations were confirmed using Sanger sequencing (see Methods).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



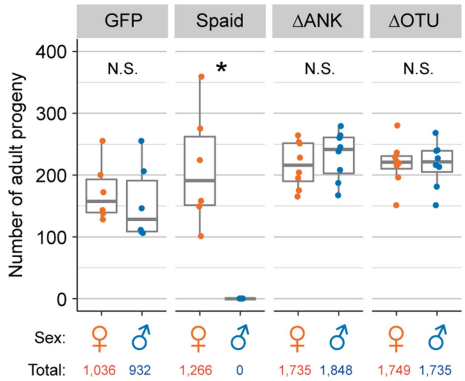
Extended Data Fig. 4 | Neural defects of Spaid-expressing embryos. Representative images of stage 13–14 female (**a**, $n = 14$) and male (**b**, $n = 16$) embryos maternally expressing Spaid and stained for TUNEL (green) and neural cells (Elav, magenta). Single-channel images of Elav are also shown. The outlined region in **b** is magnified in **c** with single-channel images of Elav and TUNEL. Embryos were co-stained for Elav, TUNEL, Sxl and DNA, and selected channels are shown in **a–c** and Fig. 2a, b, respectively.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



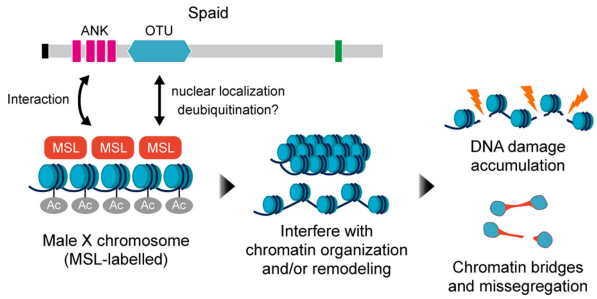
Extended Data Fig. 5 | Spaid acts through the dosage compensation machinery. a–c, Representative images of stage 13–14 embryos ectopically expressing the MSL complex (*H83M2* transgene), stained for TUNEL (green) and Sxl (magenta). GFP-expressing control female (**a**, $n = 15$), Spaid-expressing female (**b**, $n = 20$), and male (**c**, $n = 19$) embryos are shown. **d**, Quantification of TUNEL signals in *H83M2* embryos at stages 13–14. Different characters indicate significantly different groups ($P < 0.001$; Steel–Dwass test; see Supplementary Table 2). The box and dot plot (females, red; males, blue) is as in Fig. 1c and sample sizes (n, number of embryos) are shown at the bottom. **e**, **f**, Representative images of epithelial cells of stage 8–10 male embryos expressing GFP (**e**, $n = 25$) and Spaid (**f**, $n = 25$), stained for DNA (green) and MSL1 (magenta) from the datasets analysed in Fig. 3. All *UAS* transgenes were expressed maternally.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 6 | Expression of Spaid using a weak *GAL4* driver. The number of adult progeny (females, red; males, blue) obtained from crosses between the *armadillo-GAL4* driver line (weak and ubiquitous expression) and four *UAS* transgenic lines (GFP, Spaid, ΔANK, and ΔOTU; $n = 6$ independent crosses for GFP and Spaid, $n = 8$ independent crosses for ΔANK and ΔOTU). The *UAS-GFP* line was used as a negative control. With this weak *GAL4* driver, Spaid still eliminated all male progeny, while both ΔANK and ΔOTU had no impact on male viability. An asterisk indicates the statistically significant difference ($P < 0.01$; N.S., not significant, $P > 0.05$; two-tailed Mann–Whitney U test; see Supplementary Table 2). Box and dot plots are as in Fig. 1c. The total numbers of adult counts for each genotype and sex are shown at the bottom.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | A proposed model for Spaid-induced male-killing phenotypes. Spaid utilizes the OTU domain and ankyrin repeats (ANK) to target the host nucleus and the male X chromosome. 'MSL' and 'Ac' indicate the dosage compensation complex and resultant histone acetylation, respectively. See text for other explanations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0080-8

Chromatin analysis in human early development reveals epigenetic transition during ZGA

Jingyi Wu[1,7,8,9], Jiawei Xu[2,9], Bofeng Liu[1,9], Guidong Yao[2,9], Peizhe Wang[3,9], Zili Lin[1,9], Bo Huang[4], Xuepeng Wang[5,6], Tong Li[2], Senlin Shi[2], Nan Zhang[2], Fuyu Duan[3], Jia Ming[3], Xiangyang Zhang[2], Wenbin Niu[2], Wenyan Song[2], Haixia Jin[2], Yihong Guo[2], Shanjun Dai[2], Linli Hu[2], Lanlan Fang[2], Qiujun Wang[1], Yuanyuan Li[1], Wei Li[5,6], Jie Na[3]*, Wei Xie[1]* & Yingpu Sun[2]*

Upon fertilization, drastic chromatin reorganization occurs during preimplantation development[1]. However, the global chromatin landscape and its molecular dynamics in this period remain largely unexplored in humans. Here we investigate chromatin states in human preimplantation development using an improved assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq)[2]. We find widespread accessible chromatin regions in early human embryos that overlap extensively with putative *cis*-regulatory sequences and transposable elements. Integrative analyses show both conservation and divergence in regulatory circuitry between human and mouse early development, and between human pluripotency in vivo and human embryonic stem cells. In addition, we find widespread open chromatin regions before zygotic genome activation (ZGA). The accessible chromatin loci are readily found at CpG-rich promoters. Unexpectedly, many others reside in distal regions that overlap with DNA hypomethylated domains in human oocytes and are enriched for transcription factor-binding sites. A large portion of these regions then become inaccessible after ZGA in a transcription-dependent manner. Notably, such extensive chromatin reorganization during ZGA is conserved in mice and correlates with the reprogramming of the non-canonical histone mark H3K4me3, which is uniquely linked to genome silencing[3–5]. Taken together, these data not only reveal a conserved principle that underlies the chromatin transition during mammalian ZGA, but also help to advance our understanding of epigenetic reprogramming during human early development and in vitro fertilization.

We sought to investigate accessible chromatin in human preimplantation embryos using ATAC-seq[2]. Our optimized ATAC-seq protocol (miniATAC-seq; Methods) can acquire high-quality data using as few as 20 cells (Extended Data Fig. 1a–c). We collected two-cell and eight-cell human embryos, inner cell masses (ICMs) from embryonic day 5 (E5) human blastocysts, and human embryonic stem (ES) cells (Extended Data Fig. 2a). The isolated ICMs are post lineage segregation and probably include epiblast and primitive endoderm cells (Extended Data Fig. 2b). We also performed RNA sequencing (RNA-seq) analysis for each stage and validated these data (Extended Data Fig. 2c, d). We then conducted two replicates of ATAC-seq for each stage, and an additional two replicates using triploid embryos (derived from zygotes with three pronuclei, or 3PN), which are more available during in vitro fertilization (IVF). All replicates showed highly consistent results (Fig. 1a, Extended Data Fig. 3a, b and Supplementary Table 1). We identified 22,977, 40,426, 44,907 and 46,172 ATAC-seq peaks for the two-cell embryos, eight-cell embryos, ICMs and ES cells, respectively, with comparable genome coverages to that for ES cell DNase sequencing (DNase-seq; Extended Data Fig. 3c). ICM ATAC-seq peaks have the highest overlap with ES cell DNase I hypersensitive sites, and the overlap decreased in early-stage embryos (Extended Data Fig. 3d). *NANOG* is induced at the eight-cell stage, which coincides with ATAC-seq enrichment at promoters and nearby putative enhancers (Fig. 1b). *POU5F1* is slightly upregulated at the eight-cell stage and is highly activated in the ICMs and ES cells (Fig. 1b). Notably, ATAC-seq peaks near *POU5F1* transit from intronic regions to promoter upstream loci from eight-cell human embryos to ES cells, indicating a likely switch of regulatory elements between naive and primed pluripotency, as in the mouse[6]. Concordant gene expression and ATAC-seq was also found for *ZSCAN5B* (Extended Data Fig. 3e). Genome-wide, as human major ZGA starts between the four- and eight-cell stages[7], we identified stage-specifically expressed genes and examined their promoter accessibility (Extended Data Fig. 4a). We found promoters with constantly high accessibility that are preferentially CpG-rich as reported previously[8], whereas promoters with constantly low accessibility are generally CpG-poor. Interestingly, a group of genes showed dynamic promoter accessibility that correlates with gene expression. These genes preferentially function in development, differentiation and morphogenesis, as also evidenced by their histone 3 Lys27 trimethylation (H3K27me3) enrichment in ES cells and fibroblasts (Extended Data Fig. 4b, c). Therefore, promoter accessibility in early human development correlates with both gene activities and CpG densities (Extended Data Fig. 4d).

Next, we sought to investigate distal open chromatin loci. Certain repeats are highly active in mammalian preimplantation embryos[9]. In contrast to ES cells and somatic cells, distal ATAC-seq peaks preferentially enrich for repeats at the eight-cell stage, especially for Alu, SINE and long-terminal repeat (LTR) retrotransposons (Extended Data Fig. 5a). ERVK, SVA and ERV1 are enriched at both the eight-cell and ICM stages. Both ERVK and SVA are expressed in human preimplantation embryos[9]. Thus, accessible chromatin is extensively shaped by transposable elements upon human ZGA. As distal accessible chromatin frequently harbours transcription factor-binding sites, we attempted to infer key factors that may regulate transcription circuitry in early development. We first identified stage-specific distal ATAC-seq peaks (Extended Data Fig. 5b). Eight-cell specific peaks are generally present near genes involved in RNA localization and processing. By contrast, distal peaks in ES cells, but not those in the ICM, are preferentially associated with genes involved in SMAD/activin signalling and stem-cell regulation. Notably, activin A is a core signalling module that maintains primed ES cells but destabilizes naive pluripotency[10]. Interestingly, we also observed widespread open chromatin in the pre-ZGA two-cell embryos (Extended Data Fig. 5b). However, unlike in other stages, their predicted targets (Methods) generally lack transcripts at the two-cell stage. It is possible that pre-ZGA accessible elements bind transcription

[1]Center for Stem Cell Biology and Regenerative Medicine, MOE Key Laboratory of Bioinformatics, THU-PKU Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, China. [2]Center for Reproductive Medicine, Henan Key Laboratory of Reproduction and Genetics, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China. [3]Center for Stem Cell Biology and Regenerative Medicine, School of Medicine, Tsinghua University, Beijing, China. [4]PKU-THU Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China. [5]State Key Laboratory of Stem Cell and Reproductive Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. [6]University of Chinese Academy of Sciences, Beijing, China. [7]Department of Pathology and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [8]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [9]These authors contributed equally: Jingyi Wu, Jiawei Xu, Bofeng Liu, Guidong Yao, Peizhe Wang, Zili Lin. *e-mail: jie.na@tsinghua.edu.cn; xiewei121@tsinghua.edu.cn; syp2008@vip.sina.com

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

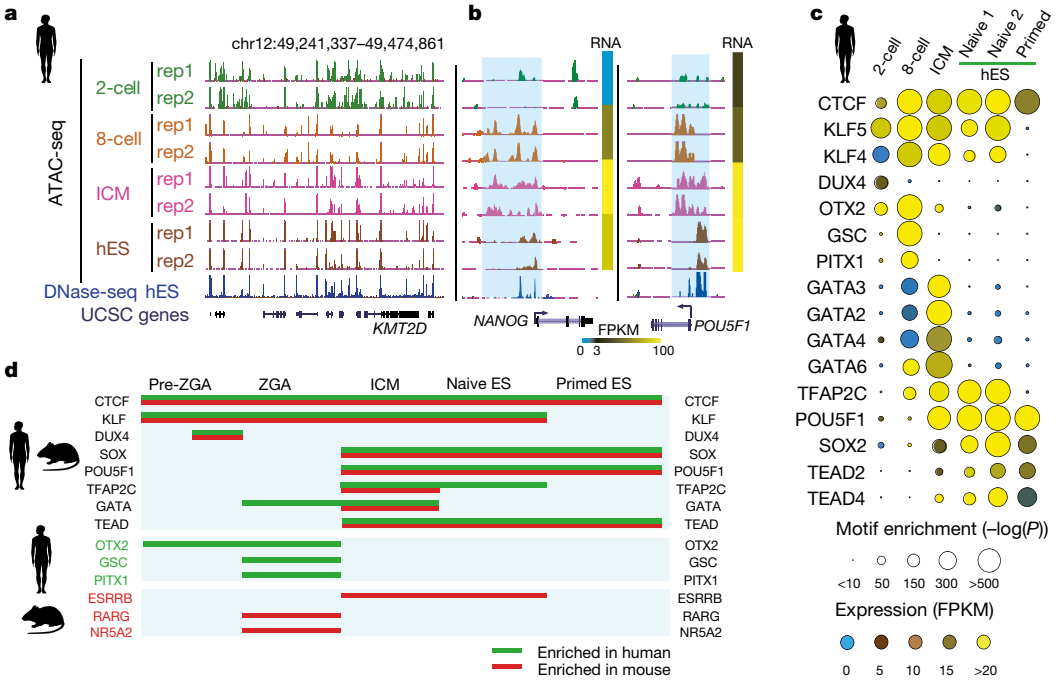


Fig. 1 | Accessible chromatin landscape in human preimplanation embryos. a, The UCSC browser view showing ATAC-seq signals in human early embryos (two biological replicates; rep 1 and rep 2). **b**, The UCSC browser views and heat maps showing ATAC-seq enrichment near representative genes and their expression levels detected by RNA-seq. Open chromatin regions are shaded. FPKM, fragments per kilobase of transcript per million mapped reads. **c**, Transcription factor motifs identified from distal ATAC-seq peaks at each stage in human embryos (two-cell, eight-cell and ICM) and human ES cells (primed, naive 1[14] and naive 2[13]). **d**, Schematic representation of putative regulatory transcription factors in early development for those shared by human and mouse as well as those specific to each species. Only enriched and expressed transcription factors are shown.

factors, but do not serve as conventional enhancers to regulate nearby genes (discussed later). Next, HOMER[11] analysis unveiled highly stage-specific transcription factor motifs in distal peaks (Fig. 1c). For example, 'maternal transcription factors' (two-cell stage and onwards) includes CTCF, KLF and OTX2. 'Minor ZGA transcription factor' (two-cell stage) includes DUX4, a key activator of early ZGA genes[12]. 'Major ZGA transcription factor' (eight-cell and onwards) includes GSC, PITX1, TFAP2C and GATA factors. 'Pluripotency transcription factors' (ICMs and ES cells) contains SOX2 and POU5F1. Notably, although conventional ES cells are typically derived from ICMs, they are generally in a primed pluripotent state[10]. However, ES cells could also be maintained in a naive state that corresponds to an earlier developmental stage[10]. Therefore, we performed ATAC-seq for naive human ES cells[13,14] (Methods). Interestingly, both KLF and TFAP2C are strongly enriched in ICMs and naive ES cells but not primed ES cells (Fig. 1c). KLF factors are known to have crucial roles in naive pluripotency[10,15]. TFAP2C is a key germline transcription factor and also a regulator for trophectoderm in mice[16]. Interestingly, *TFAP2C* is highly expressed in all lineages in human blastocysts (Extended Data Fig. 5c), suggesting human-specific functions. Finally, putative enhancer usages near *POU5F1* in naive ES cells are also more similar to the ICM than primed ES cells (Extended Data Fig. 5d). Notably, both naive and primed ES cells differ from the ICM as they lack the GATA factors that regulate the primitive endoderm lineage[17]. These data indicate that the regulatory network of naive human ES cells is indeed more similar to that of ICMs.

We then compared human and mouse transcriptional programs using previous results or ATAC-seq datasets from mouse early embryos[18], mouse ES cells (naive)[18] and mouse epiblast stem cells (EpiSCs; primed)[19]. CTCF, KLF, SOX2, POU5F1, GATA and TEAD are all well conserved between human and mouse (Fig. 1d). We also found species-specific transcription factors for preimplantation development including GSC, OTX2 and PITX1 (human) and ESRRB, NR5A2 and RARG (mouse)[18] (Fig. 1d). Notably, GSC (mesendoderm), OTX2 (neuronal lineage) and PITX1 (limb development) are all lineage regulators. KLF motifs are again enriched in ICMs and mouse ES cells but not in mouse EpiSCs. Hence, human and mouse have both conserved and divergent regulatory circuitry for early development.

We also observed widespread open chromatin at the two-cell stage (Fig. 1a). To determine whether this may be associated with minor ZGA[7], we identified genes ($n = 1{,}018$) with increased transcript levels at the two-cell stage compared to metaphase II (MII) oocytes. Unexpectedly, most of these transcripts are also present in germinal vesicle oocytes in our study and also in MII oocyte using a total RNA-seq dataset[20] (Fig. 2a, left and 2b). Importantly, the eight-cell-specific genes, but not the two-cell-specific genes, are sensitive to the transcription inhibitor α-amanitin (Extended Data Fig. 6a and Methods). Consistently, two-cell-specific genes on average only show basal levels of promoter ATAC-seq enrichment (Extended Data Fig. 6b). Shortening poly(A) tails in oocytes is thought to prevent maternal mRNA translation without triggering destabilization[21]. Thus, post-transcription regulation, such as deadenylation in oocytes and polyadenylation in early embryos, may account for a large portion of the differences between mRNA-seq data (this study and earlier studies[22,23]) and total RNA-seq[20]. After excluding maternal genes (expressed in germinal vesicle or MII oocytes), we were able to identify 75 potential minor ZGA genes activated at the two-to-four-cell stages (Extended Data Fig. 6c), and this is far fewer than the number of accessible ATAC-seq peaks we detected. However, it is worth noting that we cannot rule out the possibilities that low-input RNA-seq did not fully capture pre-ZGA transcriptional activities. We then asked whether two-cell accessible chromatin may poise genes for activation. Indeed, among all two-cell accessible promoters (maternal genes excluded) ($n = 1{,}891$), 86% are shared by the eight-cell embryos (Fig. 2c). These promoters are preferentially CpG-rich and transcribed at the eight-cell stage (Fig. 2a, right, and Fig. 2c). Conversely, eight-cell ZGA genes preferentially have accessible promoters at the two-cell stage (44% versus 29% for random genes). Notably, this is not observed for two-cell-specific open promoters (Fig. 2c and Extended Data Fig. 6d). An expanded analysis revealed that these promoters tend to reside in DNA hypomethylated regions in

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

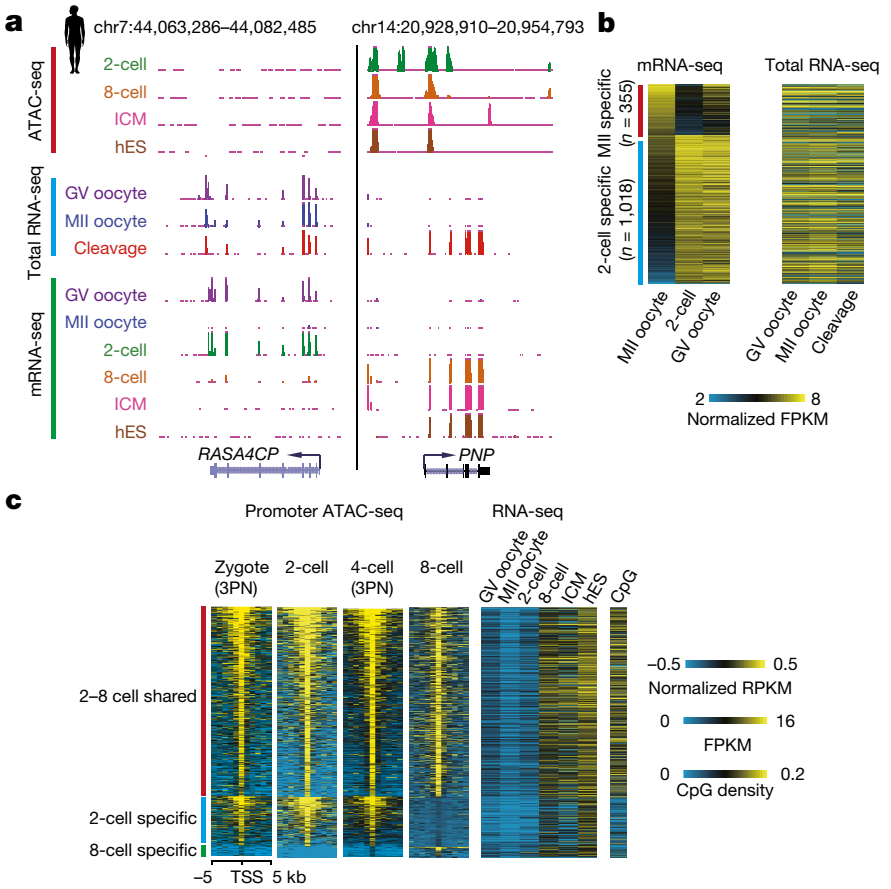


Fig. 2 | Transcription and promoter regulation in human early development. a, The UCSC browser view for genes showing transcripts but without accessible promoters (left), or vice versa (right) at the two-cell stage in humans. Total RNA-seq data were previously published[20]. **b**, Heat maps showing the expression levels (mRNA-seq) of differentially expressed genes between MII oocytes and two-cell embryos (left). Their expression level in oocyte and early embryos detected by total RNA-seq[20] are also shown (right). Cleavage stage denotes pooled two-to-eight-cell embryos. **c**, Heat maps showing ATAC-seq enrichment around accessible promoters identified at both the two- and eight-cell stages ($n = 1,444$), only at the two-cell stage ($n = 375$) or only at the eight-cell stage ($n = 82$) (left). Gene expression levels and promoter CpG density (number of CpGs per base pair) are also shown (right). Maternal genes (FPKM > 1 in either MII or germinal vesicle (GV) oocytes) were excluded to avoid confounding effects. TSS, transcription start site.

oocytes and blastocyst (despite its global hypomethylation)[24] (Extended Data Fig. 6e) (discussed later). Finally, by performing ATAC-seq on 3PN one-cell and four-cell embryos, we found that accessible chromatin at both the one- and four-cell stages highly resembles that of the two-cell stage (Fig. 2c and Extended Data Fig. 7a–c). Hence, pre-ZGA accessible chromatin is established in zygote and transits to a post-ZGA state at the eight-cell stage. In summary, these results suggest that some, but not all, pre-ZGA accessible promoters may poise genes for activation.

Interestingly, two-cell embryos have abundant distal ATAC-seq peaks ($n = 14,770$) with low CpG densities (Fig. 3a and Extended Data Fig. 8a), raising the question of whether they reflect the binding of transcription factors before ZGA[7]. In total 32% of them are shared by the eight-cell embryos and they are indeed enriched for motifs of maternal transcription factors such as CTCF, KLF and OTX2 (Extended Data Fig. 8b). The rest only occur at the two-cell stage and seem unlikely to poise chromatin for local gene activation (Extended Data Figs. 5b and 8b). These loci still enrich for enhancer marks (H3K4me1 and H3K27ac) in human ES cells and transcription factor-binding sites[25], albeit at lower levels (Fig. 3a,b and Extended Data Fig. 8c). These peaks tend to reside in DNA hypomethylated regions in human oocyte and blastocyst[24] (Fig. 3a, c). Similar patterns are found for the one- and four-cell stages, but are not for ATAC-seq peaks that are specific for other stages or methylomes of sperm or human ES cells. In human oocytes, untranscribed regions are poorly methylated[24], forming partially methylated domains (PMDs). We found that 73% of two-cell peaks are present in human oocyte PMDs (Extended Data Fig. 8d). Next, we asked whether the switch of chromatin accessibility during ZGA depends on transcription. We treated 3PN human zygotes with α-amanitin (with derived embryos term as transcription-blocked embryos, or TBEs). Both TBEs and controls developed normally to eight-cell embryos, which were collected at the late eight-cell stage for ATAC-seq analyses. We found that ATAC-seq enrichment in TBEs is highly similar to pre-ZGA embryos but is distinct from post-ZGA embryos (Fig. 3a, d). Distal open chromatin of TBEs, but not controls, is strongly enriched in oocyte PMDs (Extended Data Fig. 8e, f). Thus, the transition of accessible chromatin landscapes during human ZGA is transcription dependent.

We and others previously showed that in the mouse, oocyte PMDs are uniquely marked by non-canonical forms of histone modifications H3K4me3 and H3K27me3 (but preferentially in non-overlapping sub-regions)[3,4,26]. Depletion of H3K4me3 results in silencing defects in mature oocytes[3,5]. Non-canonical H3K4me3 (ncH3K4me3) is further inherited to mouse early embryos after fertilization, before being rapidly erased after ZGA possibly by the activated demethylases KDM5A and KDM5B[4]. After ZGA, human *KDM5B* is similarly induced (Extended Data Fig. 9a) and H3K4me3 is decreased based on immunofluorescence analysis[4,27]. Therefore, we asked whether pre-ZGA accessible chromatin may be associated with histone modifications such as ncH3K4me3. As it is challenging to perform chromatin immunoprecipitation followed by sequencing (ChIP–seq) in human embryos, we compared histone modifications in the mouse

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

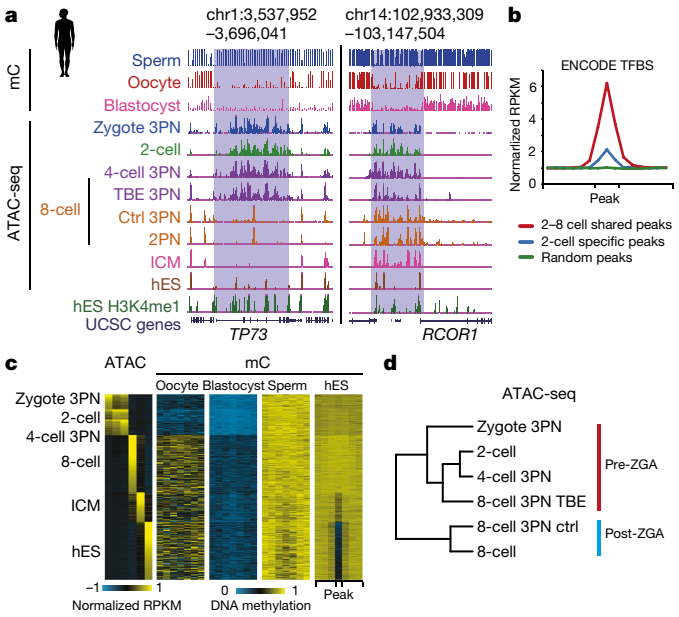


Fig. 3 | Distal accessible chromatin in human early development. a, The UCSC browser view showing DNA methylation levels (methylcytosine, or mC) in human sperm, oocyte[24], blastocyst[24], ATAC-seq enrichment in human early embryos, TBEs and ES cells, and H3K4me1 in human ES cells (ENCODE). Ctrl, control. **b**, The enrichment of ENCODE transcription factor-binding site (TFBS) density around two-cell-specific or two-to-eight-cell shared distal peaks. A random set of peaks that match the lengths of individual two-cell-specific peaks was used as a control. The upstream and downstream regions are 3 × peak lengths away from a peak boundary. **c**, Heat maps showing the oocyte, blastocyst, sperm and ES cell (ENCODE) DNA methylation levels around stage-specific distal ATAC-seq peaks (zygote $n = 889$, two-cell $n = 622$, four-cell $n = 930$, eight-cell $n = 3,382$, ICM $n = 1,777$, ES cell $n = 3,482$). The upstream and downstream regions are 2 × peak lengths away from a peak boundary. **d**, Hierarchical clustering results of whole-genome accessible chromatin states for human early embryos and TBEs.

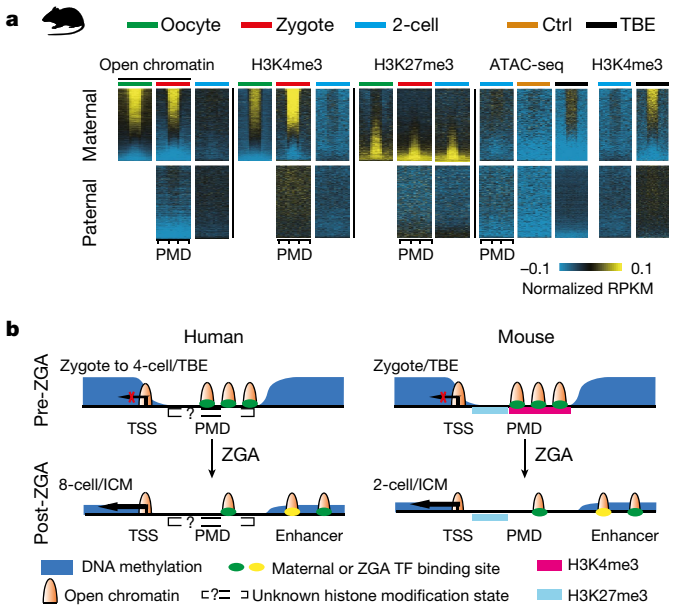


Fig. 4 | Conservation of chromatin transition during ZGA in mouse and human. a, Heat maps showing allelic open chromatin (DNase-seq (oocyte and zygote)[28] or ATAC-seq (other samples)), H3K4me3[3] and H3K27me3[26] enrichment around the oocyte PMDs ($n = 24,241$) in mouse oocyte, zygote, 2-cell embryo, TBE and its control. The upstream and downstream regions are 1 × PMD length away from the PMD boundary. **b**, A model shows the transcription and chromatin states before and after ZGA in human and mouse. In pre-ZGA embryos or TBEs, accessible chromatin occurs at CpG-rich promoters and distal regions preferentially in oocyte PMDs. After ZGA, although high-CpG accessible promoters remain open, a large fraction of pre-ZGA distal accessible regions are lost. Newly established accessible chromatin mainly occurs at putative regulatory elements such as enhancers. In mouse, dynamics of distal accessible chromatin also correlates with H3K4me3 in an allele-specific manner (not shown) before and after ZGA. Histone modification states in human are unknown.

zygote[3] to chromatin accessibility detected by DNase-seq in isolated mouse pronuclei[28]. Notably, distal open chromatin in maternal pronuclei, but not paternal pronuclei, strongly overlaps with oocyte PMDs and maternal ncH3K4me3 in mouse zygotes (Fig. 4a and Extended Data Fig. 9b). Both open chromatin and ncH3K4me3 are absent from PMDs after ZGA. Similar correlation could be extended to the paternal allele. We found the paternal H3K4me3, which is re-established after fertilization and briefly exists before ZGA[3], also co-localizes with paternal open chromatin (Extended Data Fig. 9c, d). Finally, we found that in full-grown oocytes, DNase-seq peaks also preferentially reside in oocyte PMDs and overlap with ncH3K4me3 (Fig. 4a). Thus, these data demonstrate a surprising correlation between non-canonical H3K4me3 and open chromatin in mouse mature oocytes and pre-ZGA embryos.

To determine whether the accessible chromatin transition during mouse ZGA is also transcription-dependent, we treated pronuclear stage 3 (PN3) zygotes with α-amanitin, which arrested mouse embryos at the late two-cell stage. We collected TBEs for ATAC-seq analyses when the control group developed to a later stage (eight-cells, after 45 h). On the basis of single nucleotide polymorphisms (SNPs) between the parental strains, we found the ATAC-seq enrichment in TBEs is similar to that of zygote DNase-seq in an allele-specific manner (Extended Data Fig. 10a–c). Distal open chromatin of TBEs on the maternal allele persisted in oocyte PMDs in which ncH3K4me3 is also retained after inhibition of transcription[3] (Fig. 4a). The correlation between accessible chromatin and H3K4me3 again extends to the paternal genome in TBEs (Extended Data Fig. 10a, d). Thus, the transition of both accessible chromatin and H3K4me3 during mouse ZGA depends on transcription. Interestingly, despite allele-specific accessibility, we identified very similar motifs in distal accessible chromatin on both alleles in zygotes and TBEs, mainly for maternal factors (CTCF and KLFs, such as KLF7, KLF11 and KLF17) (Extended Data Fig. 10e). As a control, transcription factors activated after ZGA (RARG, NR5A2, ESRRB, TEAD4 and GATA)[18] are not enriched, suggesting that similar maternal transcription factors interact with allele-specific chromatin landscapes before ZGA.

Using highly sensitive ATAC-seq, our data revealed both conserved and species-specific transcriptional regulatory networks in human and mouse early embryos. An unusual pre-ZGA chromatin state exists in both human and mouse, featured by extensive accessible chromatin despite the lack of transcription (Fig. 4b). Although accessible promoters preferentially enrich for CpGs, distal accessible chromatin strongly overlaps with oocyte PMDs and, in the case of mouse, ncH3K4me3. Transcription is crucial for the transition of H3K4me3 and chromatin accessibility upon ZGA. One critical question is whether pre-ZGA epigenomes in turn regulate transcription or the genome silencing. The accessible regions shared between pre- and post-ZGA stages may have a role in poised activation. By contrast, repressive marks can prevent faulty transcription during drastic epigenetic reprogramming. For instance, oocyte-deposited H3K27me3 is suggested to restrict enhancer function in fly early embryos[29] and regulate X-chromosome inactivation and allelic expression in mouse embryos[28,30]. Interestingly, maternal H3K4me3 also facilitates genome silencing in oocytes[3–5]. It is possible that ncH3K4me3 may work as sponges to facilitate docking or sequestering transcription resources when the genome is silenced. Alternatively, accessible chromatin may be created by unknown factors as 'chromatin harbours' that allow the deposition of ncH3K4me3 and docking of transcription factors. Future studies are warranted to test

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

these possibilities. Taken together, our study not only presents a global view for dynamic chromatin landscape in human embryos, but also revealed epigenomic transition during ZGA that are likely to be conserved between human and mouse.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0080-8.

Received: 27 July 2017; Accepted: 26 March 2018;
Published online 2 May 2018.

1. Burton, A. & Torres-Padilla, M. E. Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. *Nat. Rev. Mol. Cell Biol.* **15**, 723–735 (2014).
2. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
3. Zhang, B. et al. Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature* **537**, 553–557 (2016).
4. Dahl, J. A. et al. Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature* **537**, 548–552 (2016).
5. Andreu-Vieyra, C. V. et al. MLL2 is required in oocytes for bulk histone 3 lysine 4 trimethylation and transcriptional silencing. *PLoS Biol.* **8**, https://doi.org/10.1371/journal.pbio.1000453 (2010).
6. Yeom, Y. I. et al. Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. *Development* **122**, 881–894 (1996).
7. Lee, M. T., Bonneau, A. R. & Giraldez, A. J. Zygotic genome activation during the maternal-to-zygotic transition. *Annu. Rev. Cell Dev. Biol.* **30**, 581–613 (2014).
8. Fenouil, R. et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res.* **22**, 2399–2408 (2012).
9. Gerdes, P., Richardson, S. R., Mager, D. L. & Faulkner, G. J. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. *Genome Biol.* **17**, 100 (2016).
10. Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat. Rev. Mol. Cell Biol.* **17**, 155–169 (2016).
11. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
12. Iturbide, A. & Torres-Padilla, M. E. Starting embryonic transcription for the first time. *Nat. Genet.* **49**, 820–821 (2017).
13. Theunissen, T. W. et al. Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15**, 471–487 (2014).
14. Takashima, Y. et al. Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* **158**, 1254–1269 (2014).
15. Guo, G. et al. Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development* **136**, 1063–1069 (2009).
16. Kuckenberg, P., Kubaczka, C. & Schorle, H. The role of transcription factor Tcfap2c/TFAP2C in trophectoderm development. *Reprod. Biomed. Online* **25**, 12–20 (2012).
17. Morrisey, E. E. et al. GATA6 regulates HNF4 and is required for differentiation of visceral endoderm in the mouse embryo. *Genes Dev.* **12**, 3579–3590 (1998).
18. Wu, J. et al. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652–657 (2016).
19. Neijts, R. et al. Polarized regulatory landscape and Wnt responsiveness underlie Hox activation in embryos. *Genes Dev.* **30**, 1937–1942 (2016).
20. Hendrickson, P. G. et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).
21. Paynton, B. V. & Bachvarova, R. Polyadenylation and deadenylation of maternal mRNAs during oocyte growth and maturation in the mouse. *Mol. Reprod. Dev.* **37**, 172–180 (1994).
22. Xue, Z. et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597 (2013).
23. Yan, L. et al. Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
24. Okae, H. et al. Genome-wide analysis of DNA methylation dynamics during early human development. *PLoS Genet.* **10**, e1004868 (2014).
25. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
26. Zheng, H. et al. Resetting epigenetic memory by reprogramming of histone modifications in mammals. *Mol. Cell* **63**, 1066–1079 (2016).
27. Zhang, A. et al. Dynamic changes of histone H3 trimethylated at positions K4 and K27 in human oocytes and preimplantation embryos. *Fertil. Steril.* **98**, 1009–1016 (2012).
28. Inoue, A., Jiang, L., Lu, F., Suzuki, T. & Zhang, Y. Maternal H3K27me3 controls DNA methylation-independent imprinting. *Nature* **547**, 419–424 (2017).
29. Zenk, F. et al. Germ line-inherited H3K27me3 restricts enhancer function during maternal-to-zygotic transition. *Science* **357**, 212–216 (2017).
30. Inoue, A., Jiang, L., Lu, F. & Zhang, Y. Genomic imprinting of *Xist* by maternal H3K27me3. *Genes Dev.* **31**, 1927–1932 (2017).

Acknowledgements We appreciate the laboratory members' comments during preparation of the manuscript. We are grateful for the animal core facility, the sequencing core facility, and biocomputing facility at Tsinghua University. This work was supported by the National Key R&D Program of China (2016YFC0900300 to W.X. and Y.S., 2017YFA0102802 to J.N.), the National Basic Research Program of China (2015CB856201 to W.X.), the National Natural Science Foundation of China (31422031 and 31725018 to W.X., 31471404 to Y.S., 91740115 to J.N., and 31501205 to J.X.), and the THU-PKU Center for Life Sciences (W.X.). W.X. is a recipient of HHMI International Research Scholar.

Reviewer information *Nature* thanks I. Hyun and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.S. and W.X. conceived and designed the project. J.W. and B.L. developed miniATAC–seq. B.L., J.W. and J.X. performed the ATAC–seq library construction and sequencing, J.X., W.N. and N.Z. performed RNA-seq library construction and sequencing. J.W. and B.L. analysed the data. G.Y., S.S., S.D. T.L. and X.Z. collected the human oocytes, sperm and 3PN embryos, and performed ICSI and ICM separation. J.X., S.S. and X.Z. performed α-amanitin treatment experiments. Y.G., L.H., W.S., H.J. and L.F. recruited the oocyte and sperm volunteers. P.W., Z.L., B.H. and J.M. performed the mouse embryo experiments. F.D. and X.W. provided primed and naive human ES cells. Q.W. and Y.L. performed NGS sequencing. W.L., J.N., Y.S. and W.X. supervised the project or various experiments. J.W., J.X., B.L. and W.X. wrote the manuscript with the help from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0080-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0080-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.N. or W.X. or Y.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size, and investigators were not blinded to allocation during experiments and outcome assessment.

Ethics statement. This study was approved by the Institutional Review Board (IRB) of The First Affiliated Hospital of Zhengzhou University (2015KY-NO.31) and Tsinghua University (20170009), China. In accordance with the measures of the People's Republic of China on the administration of Human Assisted Reproductive Technology, the ethical principles of the Human Assisted Reproductive Technology and the Human Sperm Bank as well as the Helsinki declaration. The research followed the guiding principles of the Human Embryonic Stem Cell Ethics issued by the MOST and MOH and was regularly reviewed by the Medical Ethics Committee of The First Affiliated Hospital of Zhengzhou University. All gametes and embryos were collected voluntarily after obtaining written informed consent signed by donor couples at the Center for Reproductive Medicine in The First Affiliated Hospital of Zhengzhou University. Oocytes were obtained from the female tubal factors of infertility volunteers who have already had one or two healthy children from natural pregnancy or IVF. The eight-cell and blastocyst stages embryos were donated by couples who already had a healthy baby through IVF, with written informed consents. In the process, couples were informed that their gametes and embryos would be used to study the chromatin state and its regulation during human embryo development, and that their donation would not affect their IVF cycle.

Human gamete and early embryo collection. All gametes and embryos were collected from volunteers who were between 25 and 30 years old, with normal chromosome karyotypes and no family heredity case history, contagion case and smoke history. Healthy sperm were cryopreserved in liquid nitrogen using cryoprotectants. SpermGrad (Vitrolife) was used for gradient sperm separation. For oocytes, the controlled ovarian stimulation were carried out using GnRH analogues combined with human menopausal gonadotrophins or recombinant follicle stimulating hormone (FSH) for pituitary desensitization. Transvaginal ultrasound-guided oocyte collection was scheduled 36 h after hCG administration. Intra cytoplasmic sperm injection (ICSI) was used in this investigation rather than classical IVF to prevent contamination of sperm and cumulus cells. Fertilization was assessed 17–20 h after ICSI and embryo cleavage was recorded every 24 h. Two-cell stages embryos were frozen for storage and sample pooling, before being thawed and subjected to ATAC-seq and RNA-seq. All two-cell and eight-cell embryos and blastocysts used in this research have normal development rates and morphology. The zona pellucida of all embryos used in this study was mechanically removed to avoid blocking the reaction of Tn5 and the potential contamination of cumulus cells. The ICMs of day 5 blastocysts were isolated mechanically according to a method previously reported[31]. In brief, the biopsy pipette was inserted into the blastocoel, and the ICM cells were sucked in and gently pulled out with laser assistance. Numbers of embryos used for this study (2PN) are: two-cell rep1: 15; two-cell rep2: 17; eight-cell rep1: 8; eight-cell rep2: 10; ICM rep1: 3; ICM rep2: 3. For the 3PN embryos, the 3PN zygotes were identified after ICSI or IVF, which were collected at the one-cell stage or further cultured to the two-, four- or eight-cell stage and blastocyst using G-1 (Vitrolife) human embryos culture medium. G-2 (Vitrolife) medium was used from the eight-cell to blastocyst stage as normal 2PN embryos. Only the high-quality 3PN embryos and blastocysts without development arrest were collected.

Treatment of human embryos with α-amanitin. The 3PN zygotes were identified and collected after IVF or ICSI. About 50 3PN zygotes were thawed and transferred to G1 media in the presence or absence of α-amanitin (25 ng μl^{-1}, Sigma-Aldrich). After the control and α-amanitin-treated embryos reaching the eight-cell stage, high-quality embryos were collected and those with fragments or arrested at other stages were discarded. The zona pellucida was gently removed by laser. The cells were then washed three times with PBS to avoid any potential contamination then further subjected to ATAC-seq library preparation.

Mouse embryo collection. All animal experiments were approved by Institutional Animal Care and Use Committee (IACUC) of Tsinghua University, Beijing, China. Pre-implantation embryos were collected from 5–6-week-old C57BL/6N females (Vital River) mated with PWK/PhJ males (Jackson Laboratory). The embryo collection was performed as described previously[18]. To inhibit transcription in early embryos, PN3 zygotes were cultured in CZB supplemented in the absence or presence of α-amanitin (125 μg ml^{-1}) for about 45 h. The resulting embryos were subjected to ATAC-seq analyses.

Cell culture of human ES cells. For primed human ES cell culturing, ES cells (H1) were maintained on Matrigel (BD Biosciences)-coated plates (Corning) in E8 medium (STEMCELL Technologies) in a feeder-free and serum-free manner. Naive-state human ES cells were generated as described previously with some modifications[13,14]. In brief, for t2iL/Gö-human ES cell derivation[14], the H7 human ES cells were seeded on matrigel (BD Biosciences) coated plate 48 h before virus infection. Subsequently, cell culture medium was changed to fresh human ES cell culture medium supplemented with 3 virus concentrates containing FUW-human NANOG, FUW-human KLF2 and FUW-M2rtTA in the presence of 8 μg ml^{-1} polybrene for 24 h. Two days later the human ES cells were passaged using Accutase (Thermo Fisher) and seeded on mouse embryonic fibroblast (MEF) feeder layers in human ES cell culture medium with 10 μM Y27632. After 24 h, medium was changed to t2iL/Gö medium for further culture. After 7–14 days of culture, colonies with naive mouse ES cell-like morphology appeared and were picked to derive the t2iL/Gö-human ES cells with t2iL/Gö medium. Note that the t2iL/Gö-human ES cell proliferation rate at the initiation stage of the conversion was very low, and initial passage of t2iL/Gö-human ES cells were pursued every 6–7 days using Accutase. The 5iL/A-human ES cell derivation[13] was similar to the t2iL/Gö-human ES cell derivation, but after virus transfection, the transfected ES cells were directly cultured in 5iL/A medium supplemented with 0.5% KOSR, 8 ng ml^{-1} FGF2 and 2 μg ml^{-1} DOX. After 2 weeks of culture, colonies with naive mouse ES cell-like morphology could be isolated and expanded in 5iL/A medium without KOSR, FGF2 and DOX. All cell culture was performed under conditions of 5% oxygen at 37 °C.

The conventional H7 human ES cells were cultured on mitomycin C-inactivated MEF feeder layers in human ES cell culture medium containing 20% knockout serum replacement (KOSR) (Thermo Fisher), 78% DMEM/F12 medium (Thermo Fisher), 1% NEAA (Thermo Fisher), 2 mM GlutaMax (Thermo Fisher), 0.1 mM β-mercaptoethanol (Thermo Fisher), and 8 ng ml^{-1} FGF2 (R&D). The t2iL/Gö medium contains DMEM/DF12 and Neurobasal medium mixed at 1:1 ratio, 0.5 × N2 supplement, 0.5 × B27 supplement, 2 mM GlutaMax, 1% NEAA, 0.1 mM β-mercaptoethanol, 50 μg ml^{-1} BSA, 10 μg ml^{-1} insulin, 1 μM PD0325901, 1 μM CHIR99021, 50 μg ml^{-1} L-ascorbic acid (Sigma-Aldrich), 10 ng ml^{-1} human LIF, 2.5 μM Gö6983 (Sigma-Aldrich) and 10 μM Y27632 (Selleck). The 5iL/A medium contains DMEM/DF12 and Neurobasal medium mixed at 1:1 ratio, 0.5 × N2 supplement, 0.5 × B27 supplement, 1 mM GlutaMax, 1% NEAA, 0.1 mM β-mercaptoethanol, 50 μg ml^{-1} BSA, 1 μM PD0325901, 20 ng ml^{-1} human LIF, 1 μM IM-12 (Enzo), 0.5 μM SB590885 (R&D), 1 μM WH-4-023 (A Chemtek), 10 μM Y27632 and 20 ng ml^{-1} activin A (Peprotech).

Plasmid cloning and viral preparation. For construction of lentiviral vectors containing the human *NANOG* and *KLF2* genes, the human *NANOG* and *KLF2* coding sequences were cloned from human ES and fibroblast cell cDNA libraries, and inserted after the TetO/CMV promoter of the FUW lentivirus vector for DOX-inducible expression, respectively. Lentiviral particles containing FUW-human NANOG, FUW-human KLF2 and FUW-M2rtTA (Addgene ID: 20342) were packaged in 293FT cells with a VSVG coat. Viral supernatants were collected at 48 h after transfection, and then were filtered using a 0.45-μm filter and concentrated by high-speed centrifugation.

miniATAC-seq library generation and sequencing. The miniATAC-seq libraries were prepared as previously described with further modifications primarily for DNA purification[2,18]. In brief, samples were transferred into 6 μl lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$ and 0.5% NP-40) for 10 min on ice. After lysis, 4 μl ddH_2O, 4 μl 5 × TTBL, 5 μl TTE mix V5 (TD502, Vazyme) were added to the sample, which was mixed and incubated at 37 °C for 30 min. The reaction was terminated by incubating with 5 μl 5 × TS stop buffer (TD502, Vazyme) at the room temperature for 5 min, before 2 μl carrier RNA (20 ng μl^{-1} after 50 × dilution) (EpiTect Fast DNA Bisulphite Kit 59824) and 103 μl Tris-EDTA (TE) buffer was added to the sample. The sample was then transferred to a 1.5-ml tube. After adding 130 μl phenol–chloroform (vortexed and incubated at room temperature for 3 min), the sample was transferred to a phase-lock tube (WM5-2302820 TIANGEN). After spinning at maximum speed for 15 min, the supernatant was transferred to a new 1.5-ml tube, and 650 μl ethanol, 24 μl sodium acetate (3 M) and 2 μl glycogen were added for DNA precipitation at −20 °C overnight. The next day, DNA pellet was spun down at maximum speed for 15 min at 4 °C, washed with 75% ethanol, air dried, and resuspended in 29 μl H_2O. The phenol–chloroform purification with carrier RNA improves the efficiency for both DNA purification and removal of residual Tn5 transposase, which may otherwise interfere with subsequent reactions. DNA was transferred to a 0.2-ml PCR tube, and 5 μl N5XX primer, 5 μl N7XX primer (TD202, Vazyme), 10 μl 5 × TAB and 1 μl TAE (TD502, Vazyme) were added. PCR was performed to amplify the library for 18 cycles for about 30 embryos or ES cells using the following PCR conditions: 72 °C for 3 min; 98 °C for 30 s; and thermocycling at 98 °C for 15 s, 60 °C for 30 s and 72 °C for 3 min; following by 72 °C 5 min. After the PCR reaction, libraries were purified with the 1.5 × AMPure (Beckman) beads and were subjected to next-generation sequencing.

RNA-seq library generation and sequencing. The Smart-seq2 libraries of human oocytes and early embryos were prepared using SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Clontech, 634888) also as previously described[32]. Oocytes and early embryos from the same batch were divided into two parts, with 95% cells used for ATAC-seq and the rest for RNA-seq. Oocytes and early embryos (two-cell, four-cell and eight-cell stage, and ICMs) were lysed in lysis buffer containing RNase inhibitor according to the user manual.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

The library was quantified using Qubit and Agilent 2100 before being subjected to sequencing.

ATAC-seq data processing. The single-end ATAC-seq reads were aligned to mm9 or hg19 reference genome with random chromosome cleaned by Bowtie (version 2.2.2)[33] under the parameters -t -q -N 1 -L 25. The paired-end ATAC-seq reads were aligned with the parameters: -t -q -N 1 -L 25 -X 2000 –no-mixed –no-discordant. All unmapped reads, non-uniquely mapped reads and PCR duplicates were removed. For downstream analysis, we normalized the read counts by computing the numbers of reads per kilobase of bin per million of reads sequenced (RPKM). RPKM values were averaged for each bin between replicates. To minimize the batch and cell type variation, the RPKM values were further normalized by *Z*-score transformation. To visualize the ATAC-seq signal in the UCSC genome browser, we extended each read by 250 base pairs (bp) and counted the coverage for each base. The correlation between ATAC-seq replicates was calculated as following: each read was extended 250 bp from the mapped end position and the RPKM value was generated on a 100 bp-window base. The ATAC-seq enrichment was then summed within each 2-kb window for the entire genome and was compared between replicates. Pearson correlation was calculated and was shown. Allele assignment of sequencing reads for mouse TBE embryos was conducted as described previously[18]. In brief, to assign each read to its parental origins, we examined all SNPs in the read that showed high-quality base calling (Phred score ≥ 30). For paired-end reads, SNP information from both reads in the pair was summed and used. When multiple SNPs were present in a read (or a read pair), the parental origin was determined by votes from all SNPs and the read was assigned to the allele that had at least two thirds of the total votes.

Gene expression data processing. All RNA-seq data were mapped to hg19 genome by Tophat (version 2.4.0)[34]. The gene expression level was calculated by Cufflinks (version 2.2.1)[35] based on the hg19 refFlat annotation database from the UCSC genome browser. The Spearman correlation coefficients between RNA-seq replicates were calculated. For published total RNA-seq data[20], the expression values were quantile normalized together with the mRNA-seq data in this study to allow further comparison. For naive human ES cell gene expression microarray data[13], the expression values were downloaded and averaged for each gene. The averaged values were quantile normalized together with the FPKM value in this study to allow further comparison.

DNase-seq, ChIP–seq and DNA methylation data processing. Downloaded DNase-seq and ChIP–seq data were mapped to mm9 genome by Bowtie with similar parameters as ATAC-seq data. All unmapped reads, non-uniquely mapped reads and PCR duplicates were removed. For downstream analysis, we calculated the read counts by computing RPKM on genome 100-bp bin. Downloaded DNA methylation data were mapped to mm9 or hg19 genome by BSMAP[36] with parameters: -r 0 -p 16 -w 100 -v 0.1. PCR duplicates were removed. For each CpG site, the methylation level is calculated as the total methylated counts (combining Watson and Crick strands) divided by the total counts across all reads covering this CpG.

Identification of stage-specific genes. The minor ZGA genes activated at the two-to-four-cell stages were identified by selecting those expressed at the two-cell stage (FPKM > 5) but not expressed or expressed at low levels in MII and germinal vesicle oocytes (FPKM < 5) in both mRNA-seq and total RNA-seq, with a minimal twofold gene expression change between the two- or four-cell and oocyte stages.

For stage-specific genes in eight-cell embryos, ICMs and human ES cells, a Shannon entropy-based method was used to identify stage-specific genes, as previously described[37]. Owing to the possible confounding effects from maternally inherited RNA transcripts, ZGA-only genes were analysed, which were defined as those not expressed in MII or germinal vesicle oocytes (FPKM < 1). Genes with entropy scores of less than 2 were selected as candidates for stage-specific genes. Among these genes, we selected candidates of stage-specific genes for each stage based on the following criteria: the gene is highly expressed at this stage (FPKM > 10), and such high expression cannot be observed in other stages. These genes were then reported in the final stage-specific gene lists.

Identification of ATAC-seq peaks and their genome coverages. All the ATAC-seq peaks were called by MACS v1.4[38] with the parameters–nolambda –nomodel. ATAC-seq peaks that are at least 2.5 kb away from annotated promoters from refFlat were selected as distal ATAC-seq peaks. The genome coverages of peaks from different samples were calculated by genomeCoverageBed[39] using hg19 reference genome.

The comparison between ATAC-seq peaks and known *cis*-regulatory elements. To compare the ATAC-seq peaks identified in early embryos with the annotated *cis*-regulatory elements, we calculated the overlap between the ATAC-seq peaks of different stages and annotated promoters (TSS $\pm$ 0.5 kb). Distal peaks were then compared to distal DNase I hypersensitive sites in ES cells. Random peaks were generated by selecting random regions in the genome with the sizes matching each individual ATAC-seq peak.

The comparison between ATAC-seq peaks and repetitive elements. To identify the overlap between repetitive elements and promoter or distal ATAC-seq peaks, the ATAC-seq peaks were compared with the locations of annotated repeats (RepeatMasker) downloaded from the UCSC genome browser by intersectBed[39] with default parameters. As repeats of different classes vary greatly in numbers, a random set of peaks with identical lengths of ATAC-seq peaks was used for the same analysis as a control. The numbers of observed peaks that overlap with repeats were compared to the numbers of random peaks that overlap with repeats, and a log ratio value ($\log_2$) was generated as the 'observed/expected' enrichment.

Identification of stage-specific distal ATAC-seq peaks. The two-cell-specific distal ATAC-seq peaks were identified as those with high levels of ATAC-seq enrichment at the two-cell stage (normalized RPKM > 1) but not at the eight-cell stage (normalized RPKM < 0). Those with normalized RPKM > 1 at both stages were identified as shared distal peaks between the two- and eight-cell stages.

To identify stage-specific distal ATAC-seq peaks, the distal ATAC-seq peaks from human embryos of all stages and ES cells were combined, with overlapped peaks merged. The average RPKM values were calculated for these distal ATAC-seq peaks which were further normalized by the *Z*-score normalization. A Shannon entropy-based method[37] was used to identify stage-specific distal ATAC-seq peaks. We selected those with entropy less than 2 as candidates for stage-specific distal ATAC-seq peaks. The stage-specific distal ATAC-seq peaks were further defined based on the following criteria: the distal ATAC-seq peak has high enrichment at this stage (normalized RPKM > 1), and positive enrichment (normalized RPKM > 0) at no more than two additional stages. The resulting distal ATAC peaks were then reported in the final stage-specific distal ATAC-seq peak list. The functional enrichment for genes that are near stage specific distal ATAC-seq peaks was analysed using the GREAT tool[40] by default settings.

Prediction of promoter targets of putative enhancers in distal ATAC-seq peaks. To identify the potential targeted genes for stage specific enhancer (distal peaks), we computed the averaged ATAC-seq enrichment (normalized RPKM) for all distal ATAC-seq peaks and annotated promoters (TSS $\pm$ 0.5 kb). Among genes assigned to enhancers by GREAT analysis based on distances, we further calculated the correlation between the ATAC-seq enrichment at distal ATAC peaks and these promoters across human embryos and ES cells. The promoter with a Pearson correlation coefficient above 0.8 was selected as the potential target of the enhancer.

Gene ontology analysis. The DAVID web-tool was used to identify the Gene Ontology (GO) terms using databases including Molecular Functions, Biological Functions and Cellular Components[41].

Hierarchical clustering analysis. The hierarchical clustering was performed in R by hclust() function with ATAC-seq RPKM values via Spearman correlation coefficients.

Motif, enhancer mark and transcription factor-binding sequence analyses for distal ATAC-seq peaks. To find the sequence motifs enriched in distal ATAC-seq peaks, findMotifsGenome.pl from the HOMER program[11] was used. Motifs with known match in HOMER database were selected. The ChIP–seq data for human ES cell H3K27ac and H3K4me1 marks[42] and the collection of transcription factor-binding sites[25] were downloaded from the UCSC genome browser. The average RPKM values at the distal peaks and their nearby regions were calculated. The number of transcription factor-binding sites was first binned for each 100-bp window, and the average enrichment at the distal peaks and their nearby regions was calculated.

Identification of oocyte PMDs. PMDs in oocytes were identified as described previously[3]. In brief, published human[24] and mouse[43] oocyte DNA methylation datasets were downloaded. Average oocyte DNA methylation levels were calculated for 10-kb bins of the genome. Bins with average DNA methylation levels of lower than 0.5 were selected and merged into PMDs. Promoter regions ($\pm$2.5 kb) were excluded from PMDs.

The ATAC-seq enrichment in PMD. ATAC-seq RPKM is *Z*-score normalized on the genome-wide 100-bp bin with promoter regions removed. PMDs, as well as their upstream and downstream regions, were divided into five bins separately, and the mean RPKM was calculated for each bin and shown in the heat map. The average ATAC-seq RPKM for all PMDs, as well as their upstream and downstream regions, was computed and shown.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

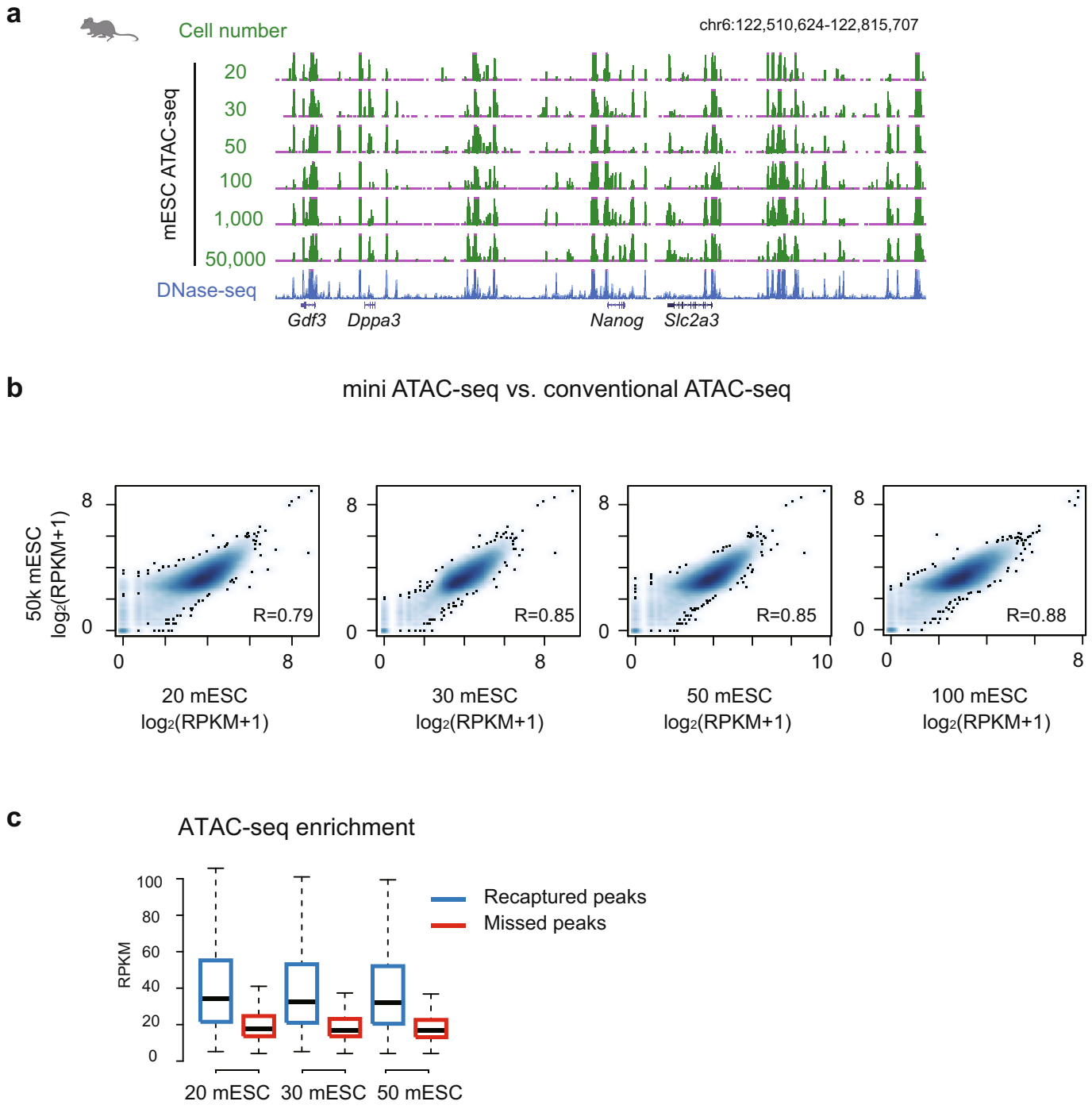
Data availability. The data sets generated and analysed during the current study are available in the Gene Expression Omnibus (GEO) with the accession number GSE101571.

31. Capalbo, A. et al. FISH reanalysis of inner cell mass and trophectoderm samples of previously array-CGH screened blastocysts shows high accuracy of diagnosis and no major diagnostic impact of mosaicism at the blastocyst stage. *Hum. Reprod.* **28**, 2298–2307 (2013).
32. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171–181 (2014).
33. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

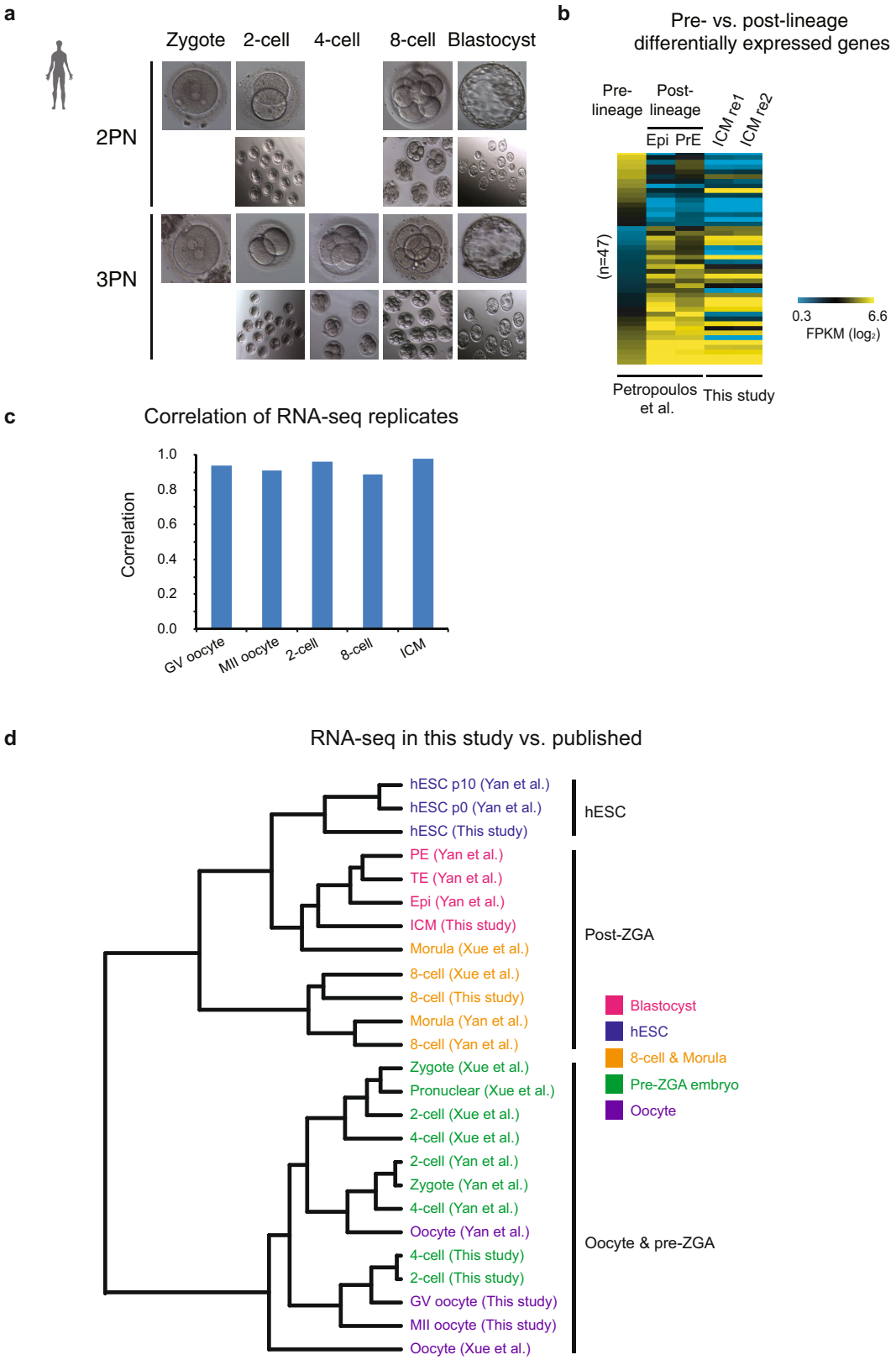
34. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
35. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7**, 562–578 (2012).
36. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
37. Schug, J. et al. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* **6**, R33 (2005).
38. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
39. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
40. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
41. Dennis, G., Jr et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, 3 (2003).
42. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
43. Kobayashi, H. et al. Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. *PLoS Genet.* **8**, e1002440 (2012).
44. Petropoulos, S. et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **167**, 285 (2016).
45. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
46. Hawkins, R. D. et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).
47. Xie, W. et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



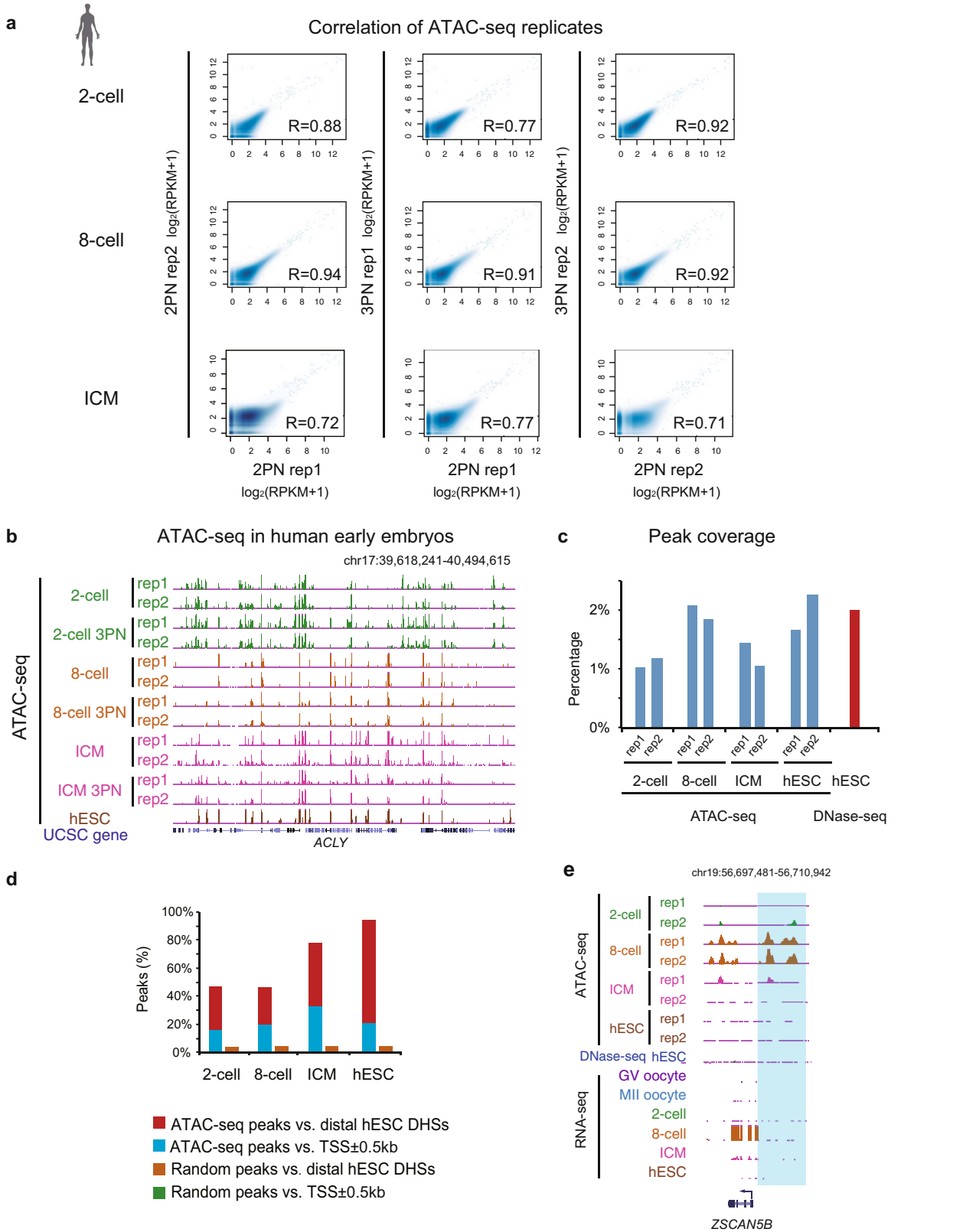
Extended Data Fig. 1 | Validation of miniATAC-seq. a, UCSC browser view showing the ATAC-seq signals using various numbers of mouse ES cells. **b**, Scatter plots comparing the miniATAC-seq signals from various numbers of mouse ES cells with conventional ATAC-seq using 50,000 mouse ES cells. The Pearson correlation coefficients are also shown. **c**, Box plots showing the ATAC-seq enrichment for peaks from 50,000 mouse ES cells that are recaptured or missed by miniATAC-seq using lower numbers of mouse ES cells.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



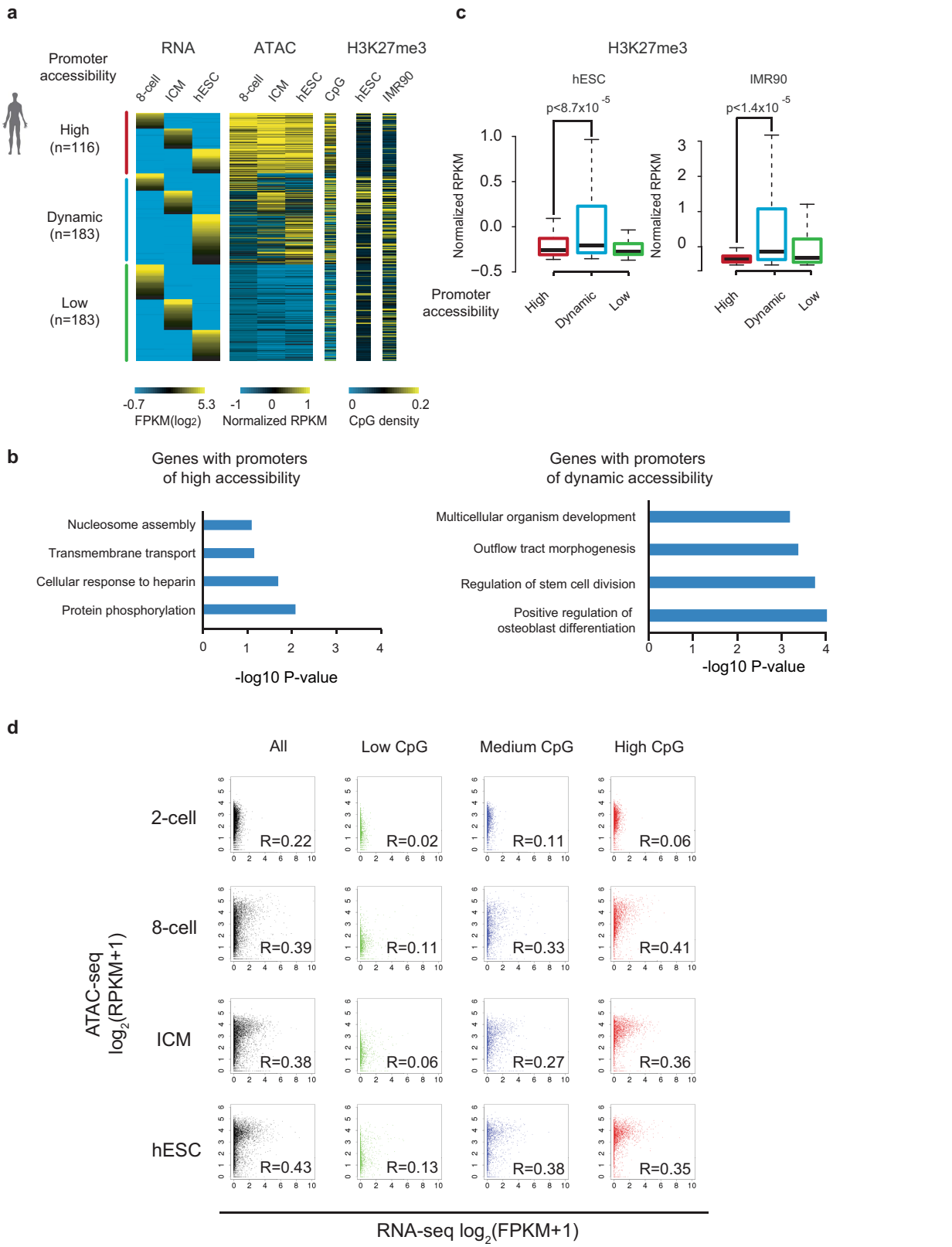
Extended Data Fig. 2 | Validation of RNA-seq data in human early embryos. a, Microscopy images of human 2PN and 3PN embryos at the zygote, two-cell, four-cell, eight-cell and blastocyst stages. ICSI was used to avoid the cumulus cell contamination. High-quality score embryos were selected for subsequent study (represented by magnified images). **b**, Heat maps showing gene expression levels for differentially expressed genes between human embryos of pre-lineage segregation and post-lineage specification[44]. Expression of ICMs in this study is also shown for two replicates. Epi, epiblast, PrE, primitive endoderm. **c**, Bar charts showing the Spearman correlation between the two replicates of RNA-seq samples. **d**, Hierarchical clustering of RNA-seq datasets from this study (using Smart-seq) and previous studies[22,23] (using a different mRNA-seq method[45]). Pearson correlation was used to measure distances. Different colours represent various stages.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



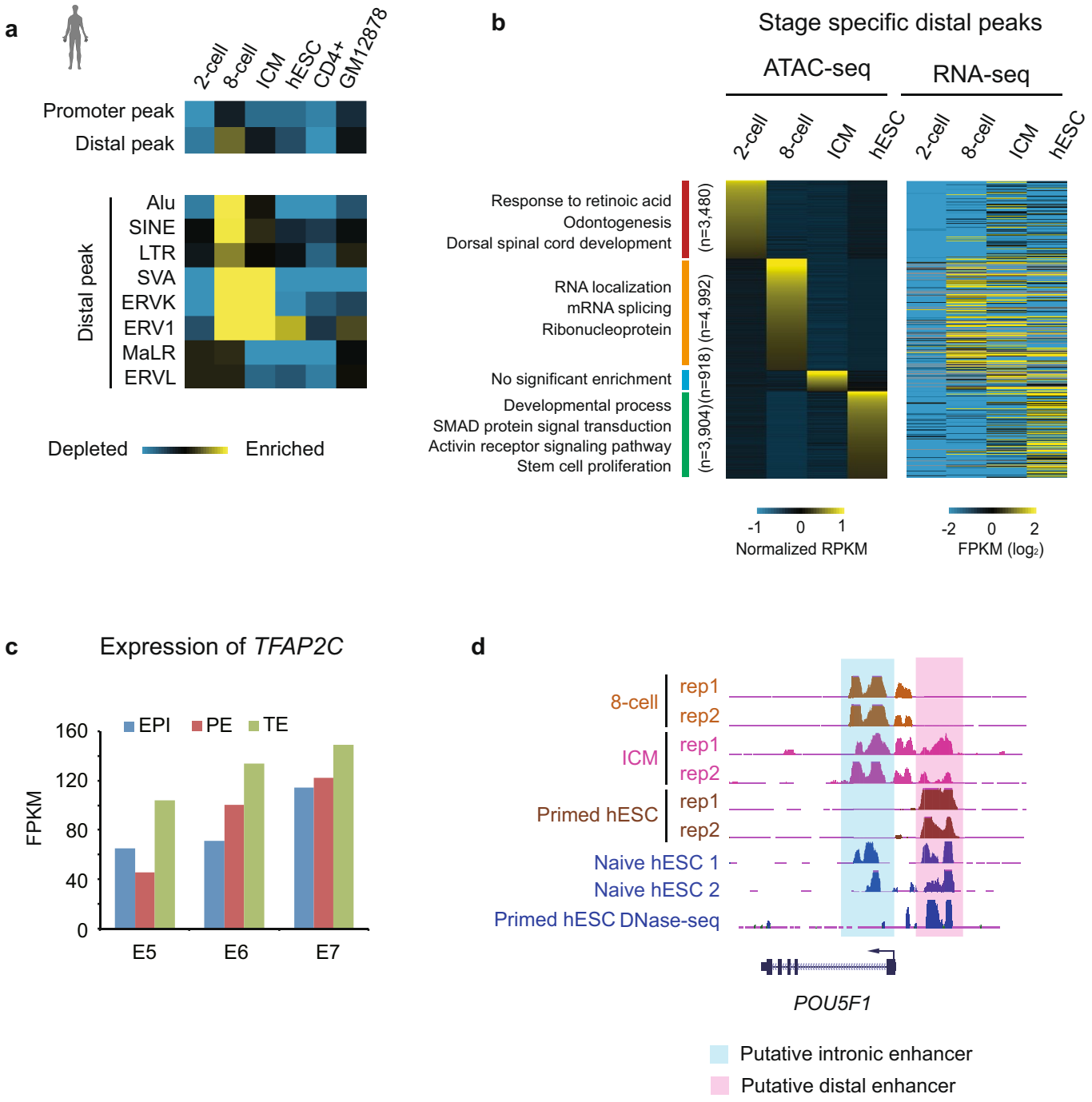
Extended Data Fig. 3 | Validation of ATAC-seq data in human early embryos. a, Scatter plots showing the ATAC-seq signals between replicates at each stage in human early development or between 2PN and 3PN embryos. **b**, UCSC browser view showing the landscape of accessible chromatin in replicates of human early embryos. **c**, Bar charts showing the genome coverages of ATAC-seq peaks of each stage or DNase-seq peaks for human ES cells (ENCODE). **d**, The overlap between ATAC-seq peaks and annotated promoters (TSS ± 0.5 kb) or distal DNase I hypersensitive sites in ES cells. A random set of peaks that match the lengths of individual ATAC-seq peaks were used as a control. **e**, UCSC browser views showing the ATAC-seq and RNA-seq enrichment near a representative gene. Open chromatin regions are shaded.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



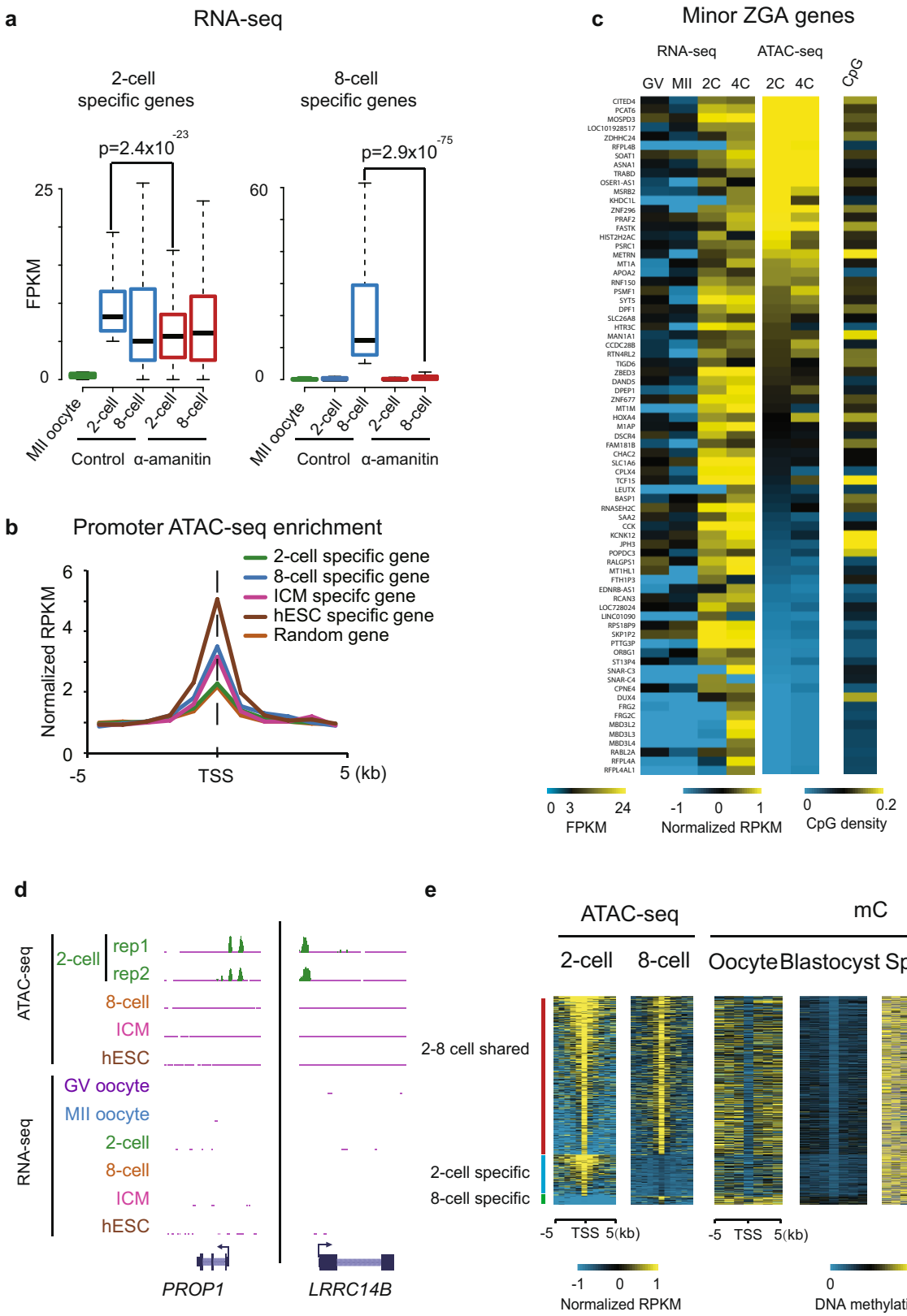
Extended Data Fig. 4 | Relationship of chromatin accessibility and transcription in human early embryos. a, Heat maps showing three classes of promoter accessibility (high, dynamic and low) for stage-specific genes (maternal genes excluded). CpG densities and H3K27me3 levels in human ES cells and fibroblasts (IMR90)[46] are also shown. **b**, GO analysis results for gene classes in **a**. (The 'low' class does not have gene set enrichment with $P < 10^{-2}$). **c**, Box plots showing promoter enrichment of H3K27me3 in human ES and IMR90 cells[46] for each class in **a**. *P* values based on a one-sided *t*-test are shown. **d**, Scatter plots showing promoter ATAC-seq enrichment and gene expression (maternal genes excluded) for all genes (black) or genes with promoters of low (green), medium (blue), or high (red) CpG densities. Spearman correlation coefficients are shown.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



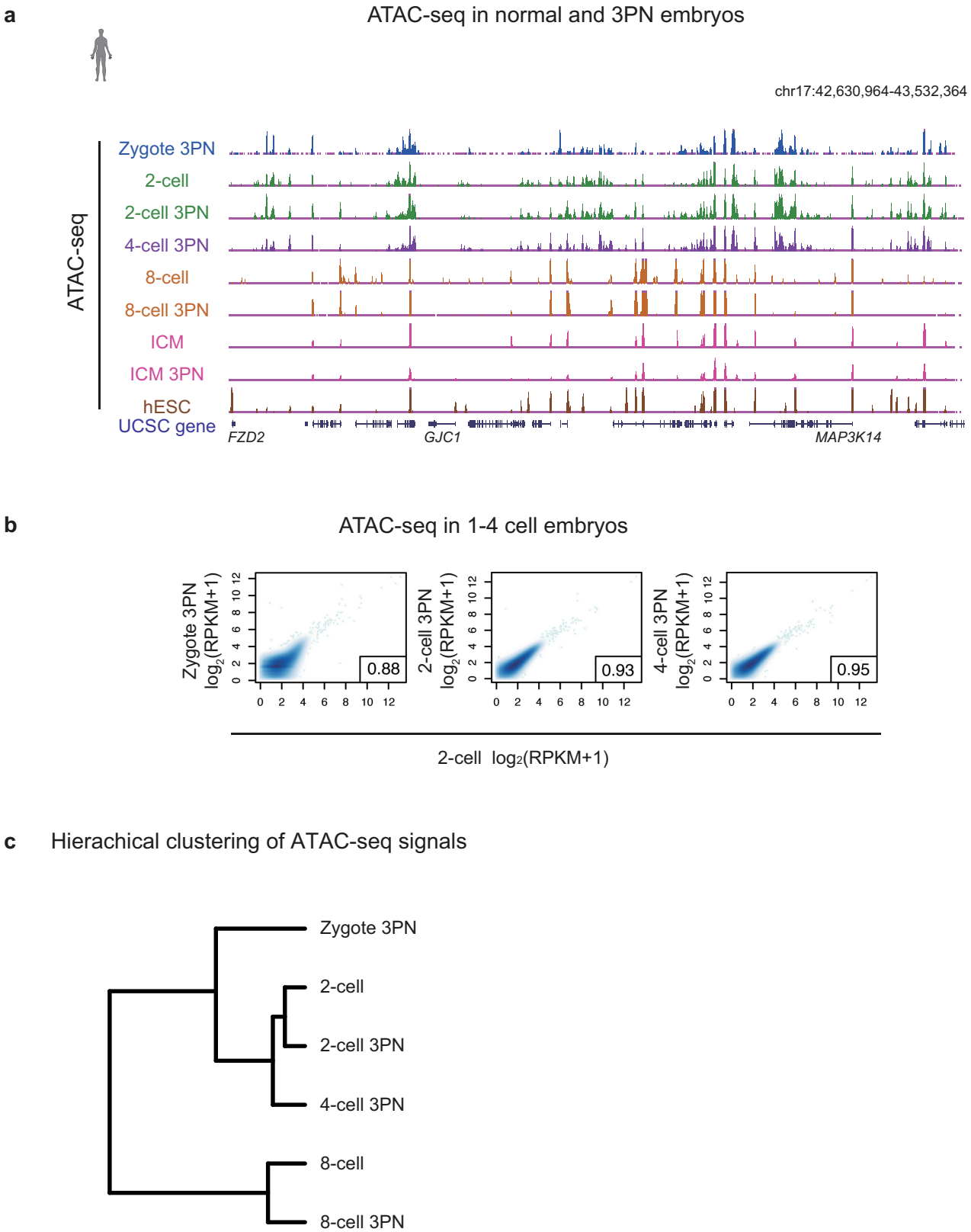
Extended Data Fig. 5 | Features of distal ATAC-seq peaks in human embryos. a, Top, enrichment of repeats in ATAC-seq promoter and distal peaks compared to that in random peaks for early human embryos, human ES cells, and somatic cell types. The enrichment was calculated as a $\log_2$ ratio for the numbers of observed peaks that overlap with repeats divided by the numbers for random peaks. A random set of peaks that match the lengths of individual ATAC-seq peaks was used. Bottom, a similar analysis was performed for the enrichment of repeat subfamilies in distal peaks. **b**, Middle, heat maps showing the ATAC-seq signals for stage-specific distal ATAC-seq peaks in human embryos. Left, GREAT analysis[40] results are also shown. Right, the expression of predicted targets among GREAT listed nearby genes for putative enhancers (distal peaks) (Methods) are shown for each stage. **c**, Bar charts showing the expression of *TFAP2C* in epiblast (EPI), primitive endoderm (PE) and trophectoderm (TE) lineages in human embryos from embryonic days 5–7 based on a previous study[44]. **d**, UCSC browser view showing the ATAC-seq signals around *POU5F1* in human early embryos and ES cells (primed, naive 1[14] and naive 2[13]). Human ES cell DNase-seq data (ENCODE) are shown as a control.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



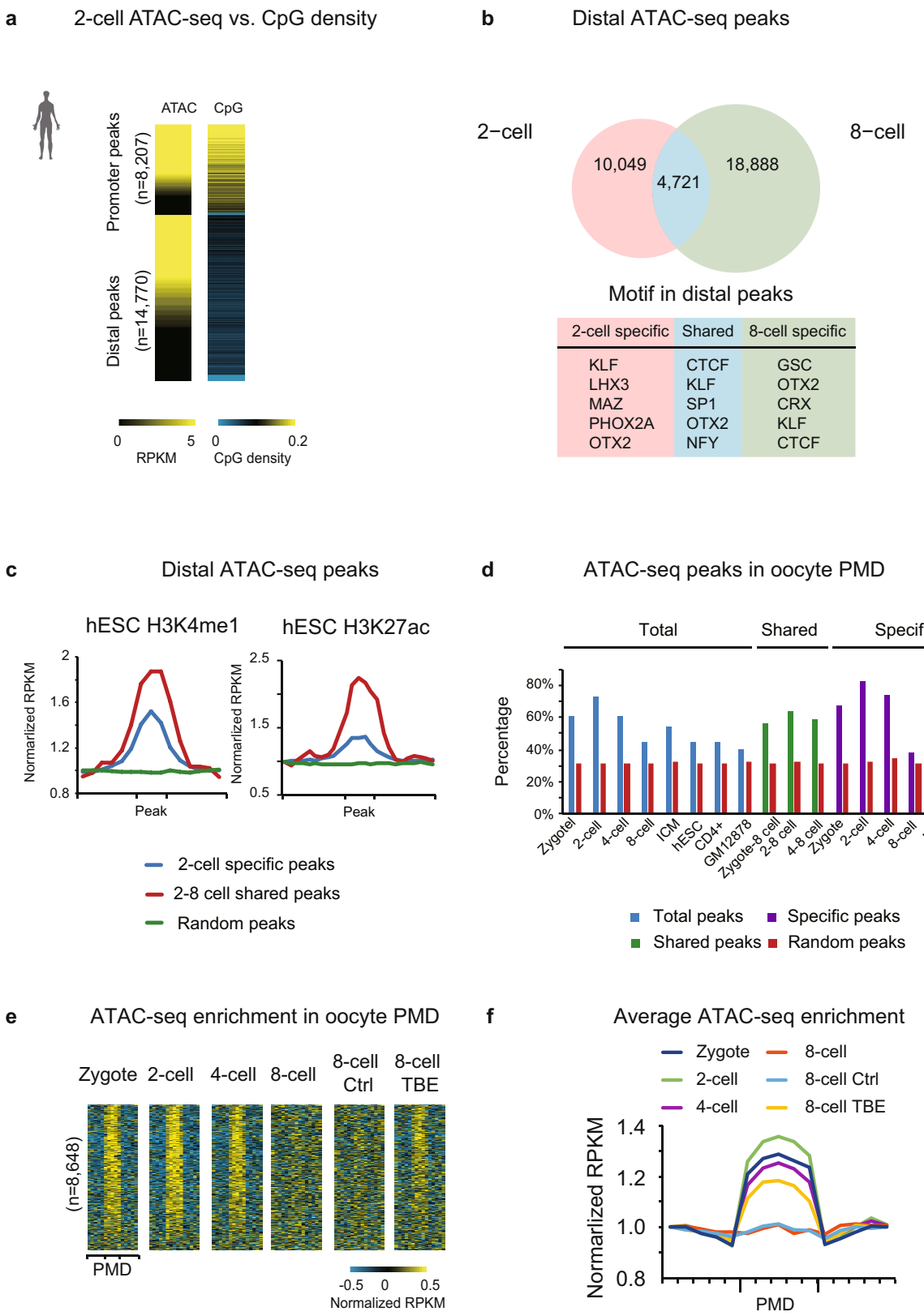
Extended Data Fig. 6 | Transcription and chromatin states before major human ZGA. a, Box plot showing the expression levels of two-cell specific (left) and eight-cell specific (right) genes in MII oocytes and embryos with or without α-amanitin treatment. *P* values based on a one-sided *t*-test are shown. **b**, The average ATAC-seq enrichment for each stage is shown at the promoters of stage specific genes at the same stage identified by mRNA-seq. Two-cell ATAC-seq enrichment for a random set of promoters were similarly analysed as a control. **c**, Heat maps showing expression levels of possible minor ZGA genes activated at the two-to-four-cell stage for their expression in germinal vesicle oocytes, MII oocytes, and two-to-four-cell embryos. Promoter ATAC-seq enrichment for two-to-four-cell stages and CpG densities are also shown. **d**, UCSC browser view showing the promoter ATAC-seq signals specifically appearing in two-cell embryos. **e**, Left, heat maps showing ATAC-seq enrichment at the accessible promoters present in both two- and eight-cell embryos, as well as those specific to each stage. Right, the human oocyte, blastocyst, sperm and ES cell DNA methylation levels around these promoters are also shown.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



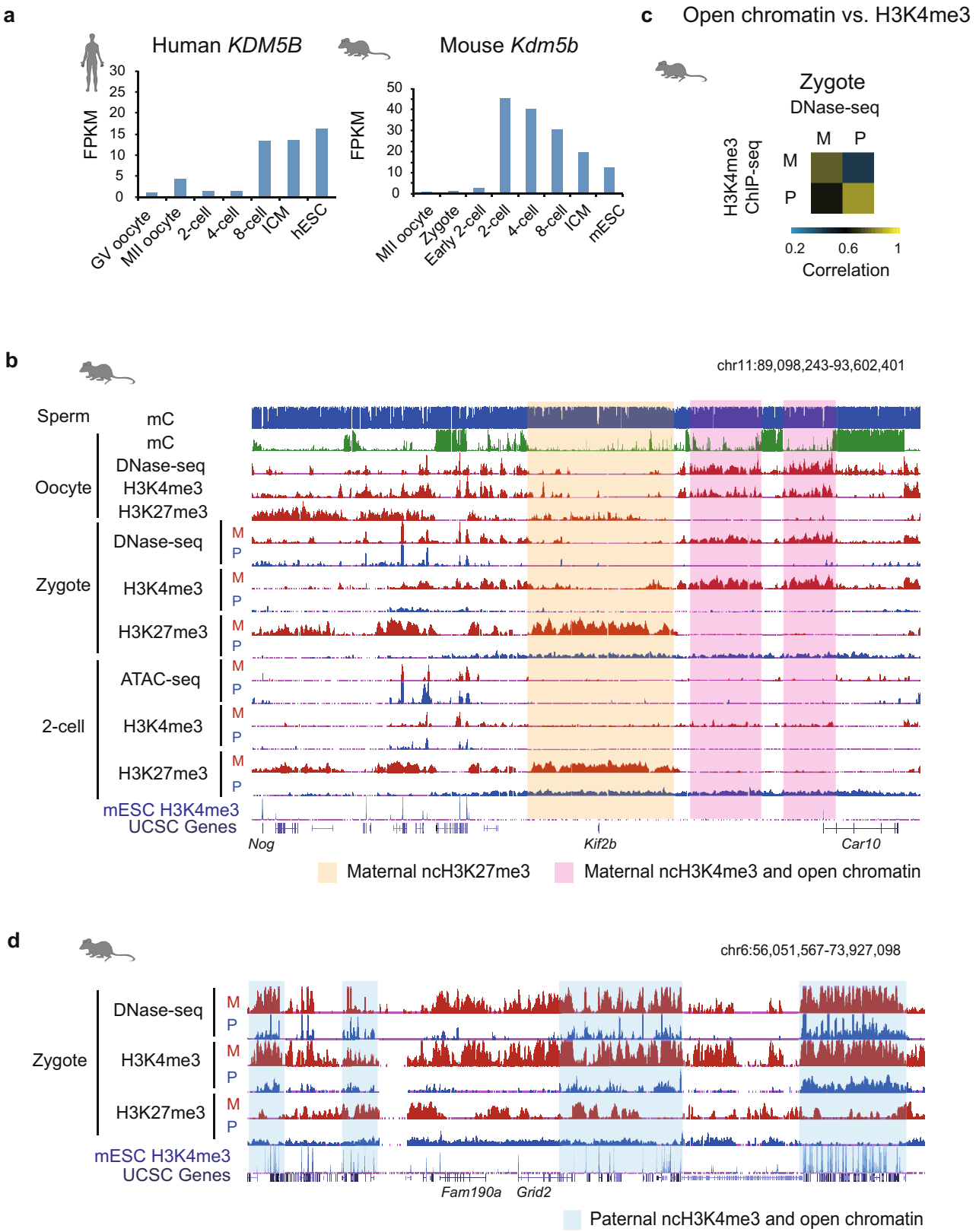
Extended Data Fig. 7 | Accessible chromatin state in one-cell and four-cell human embryos. a, UCSC browser view showing the ATAC-seq signals in human early embryos. **b**, Scatter plots showing the ATAC-seq signals between two-cell human embryos and 3PN one-cell, two-cell and four-cell embryos. **c**, Hierarchical clustering results showing the relationships of chromatin states among human embryos based on whole-genome ATAC-seq enrichment.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



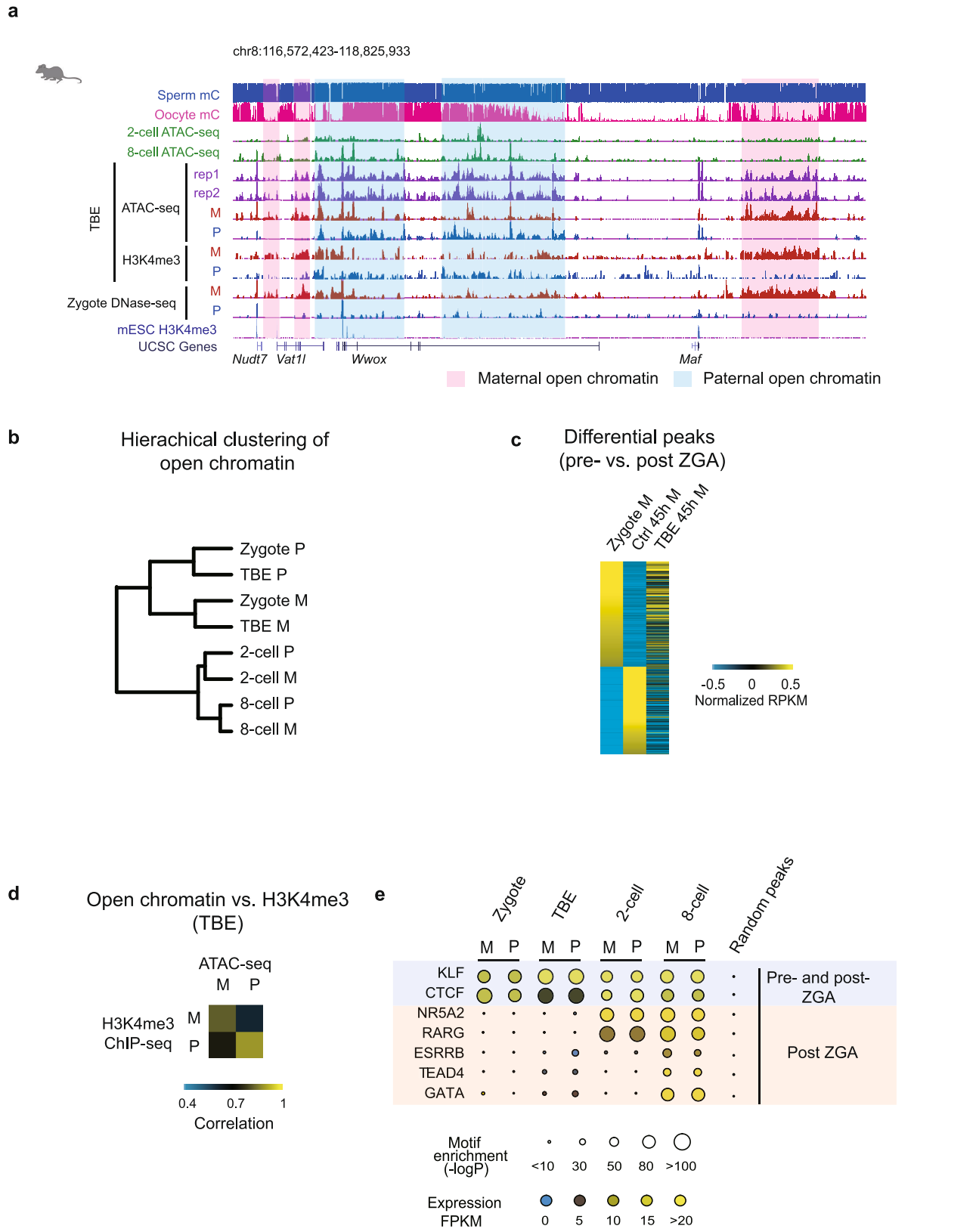
Extended Data Fig. 8 | Characterization of two-cell distal open chromatin in human embryos. a, Heat maps showing the ATAC-seq signals and CpG density for promoter and distal ATAC-seq peaks at the two-cell stage. **b**, Top, Venn diagram showing the overlap between two- and eight-cell distal ATAC-seq peaks. Bottom, motifs identified in two- and eight-cell embryo shared distal peaks as well as peaks specific for each stage are also shown. **c**, The enrichment of human ES cell H3K4me1 and H3K27ac marks[47] around two-cell-specific distal peaks or two-to-eight-cell shared distal peaks is shown. A random set of peaks that match the lengths of individual two-cell-specific ATAC-seq peaks was used as a control. The upstream and downstream regions are 2 × peak lengths away from peak boundary. **d**, Bar chart showing the percentages of ATAC-seq peaks that overlap with oocyte PMDs for total peaks, peaks shared by various stages, or peaks specific for each stage. **e**, Heat maps showing the enrichment of ATAC-seq around the human oocyte PMDs in human early embryos and TBEs. The upstream and downstream regions are 1 × PMD length away from the PMD boundary. **f**, The average enrichment of ATAC-seq around the oocyte PMDs in human early embryos and TBEs. The upstream and downstream regions are 1 × PMD length away from the PMD boundary.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | Accessible chromatin state in mouse oocytes and pre-ZGA embryos. a, The expression of human *KDM5B* and mouse *Kdm5b* in oocytes, early embryos and ES cells. **b**, UCSC browser view showing DNA methylation levels (mC) in mouse sperm and oocyte, as well as open chromatin (DNase-seq or ATAC-seq) and H3K4me3 and H3K27me3 enrichment in mouse oocytes, zygotes and two-cell embryos. Mouse ES cell H3K4me3 signals are also shown to mark the promoter regions. **c**, Heat map showing the Spearman correlation between allelic DNase-seq and H3K4me3 signals in zygotes. M, maternal; P, paternal. **d**, UCSC browser view showing allelic DNase-seq, H3K4me3 and H3K27me3 enrichment in the mouse zygote. The mouse ES cell H3K4me3 signal is also shown. Regions showing paternal open chromatin and H3K4me3 in zygotes are shaded.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Accessible chromatin state in mouse TBEs. a, UCSC browser view showing DNA methylation levels in mouse sperm and oocyte, and ATAC-seq signals in normal mouse embryos as well as allelic ATAC-seq and H3K4me3 enrichment in TBE samples. Mouse ES cell H3K4me3 signals are also shown to mark the promoter regions. **b**, Hierarchical clustering results showing the relationships of allelic accessible chromatin states between zygotes, TBEs, and two- and eight-cell embryos in mouse. **c**, Heat map showing open chromatin regions that are unique to zygotes (DNase-seq) or 45 h control embryos (ATAC-seq). The ATAC-seq enrichment in TBE samples is then matched and shown. **d**, Heat map showing the Spearman correlation between allelic ATAC-seq and H3K4me3 signals in TBEs. **e**, Transcription factor motifs identified from distal DNase-seq and allelic distal ATAC-seq peaks are shown. Motifs shared by pre-ZGA and post-ZGA stages or are specific for post-ZGA stages are noted. For transcription factors that have multiple family members with similar motifs (KLF and GATA), the highest expression and motif enrichment among all family members at each stage are shown. A random set of peaks that match the lengths and number of zygote maternal peaks was used as a control. It is worth noting that the RNA levels of CTCF appear to decline in TBEs, presumably owing to RNA degradation during extended transcription inhibition.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0081-7

Structural principles of distinct assemblies of the human α4β2 nicotinic receptor

Richard M. Walsh Jr[1,2,5], Soung-Hun Roh[3,4,5], Anant Gharpure[1,2], Claudio L. Morales-Perez[1,2], Jinfeng Teng[1,2] & Ryan E. Hibbs[1,2]*

Fast chemical communication in the nervous system is mediated by neurotransmitter-gated ion channels. The prototypical member of this class of cell surface receptors is the cation-selective nicotinic acetylcholine receptor. As with most ligand-gated ion channels, nicotinic receptors assemble as oligomers of subunits, usually as hetero-oligomers and often with variable stoichiometries[1]. This intrinsic heterogeneity in protein composition provides fine tunability in channel properties, which is essential to brain function, but frustrates structural and biophysical characterization. The α4β2 subtype of the nicotinic acetylcholine receptor is the most abundant isoform in the human brain and is the principal target in nicotine addiction. This pentameric ligand-gated ion channel assembles in two stoichiometries of α- and β-subunits (2α:3β and 3α:2β). Both assemblies are functional and have distinct biophysical properties, and an imbalance in the ratio of assemblies is linked to both nicotine addiction[2,3] and congenital epilepsy[4,5]. Here we leverage cryo-electron microscopy to obtain structures of both receptor assemblies from a single sample. Antibody fragments specific to β2 were used to 'break' symmetry during particle alignment and to obtain high-resolution reconstructions of receptors of both stoichiometries in complex with nicotine. The results reveal principles of subunit assembly and the structural basis of the distinctive biophysical and pharmacological properties of the two different stoichiometries of this receptor.

Pseudo-symmetric heteromeric proteins, exemplified by the Cys-loop receptor family, present a challenge in cryo-electron microscopy (cryo-EM), because particle alignment is dominated by low-resolution information[6,7], and at low resolution the subunit identities may not be readily distinguishable. Heterogeneity in subunit assembly adds to the challenges in particle alignment. In the nicotinic receptor, the α- and β-subunits have nearly identical secondary structural features, which makes them and their multiple assemblies indistinguishable at low resolution. We raised monoclonal antibodies against the recombinant α4β2 nicotinic receptor and isolated high-affinity antigen-binding (Fab) fragments to structurally distinguish α-subunits from β-subunits within the pseudo-symmetric pentamer. In contrast to our earlier crystallographic studies, in which we optimized expression conditions to produce more-or-less exclusively the 2α:3β assembly[8], here we intentionally produced a mixed population of both stoichiometries of the α4β2 receptor (2α:3β and 3α:2β). We co-purified the receptor–Fab complex, and prepared sample grids for cryo-EM (Extended Data Fig. 1a and Methods). Fab has little or no effect on [^{3}H]-nicotine binding and activation of the channel (Extended Data Fig. 1b–d).

Single particle cryo-EM data collection and processing revealed two distinct receptor populations, one with two Fabs bound and one with three Fabs bound, corresponding to the 3α:2β and 2α:3β stoichiometries, respectively (Extended Data Fig. 2). We obtained cryo-EM maps of the two stoichiometries to approximately 3.5 Å resolution for the receptor and approximately 4 Å resolution for the entire complex including Fab (Fig. 1a–d and Extended Data Figs. 2, 3), with map quality that was substantially improved over that of the crystallographic structure of the 2α:3β assembly[8] (Extended Data Table 1 and Extended Data Figs. 4–6). Each receptor assembly is cylinder-shaped, with a large extracellular vestibule and funnel-shaped transmembrane channel. Fabs interact exclusively with the extracellular domain (ECD) of the β2 subunits (Extended Data Fig. 1e–h). The agonist nicotine was included throughout purification, and density for it was observed at α–β and α–α interfaces in the receptor ECD. Nicotine stabilizes the ion channel in a putative closed-desensitized conformation.

We sought to understand why we did not observe more than two subunit assemblies[3,9,10], given that our structures demonstrate the occurrence of all possible interfaces between α4 and β2 subunits. Each stoichiometry contains two α–β and two β–α interfaces, plus a third unique interface type: one α–α (3α:2β) or one β–β (2α:3β) interface (Fig. 2a–d). Our reconstructions reveal that in an agonist bound, desensitized state, the α–α interface buries the most protein surface area from solvent, followed by the α–β and β–α interfaces (Fig. 2e). The β–β interface has the smallest amount of buried surface area. Calculated interfacial free energies predict preferential assembly of α–β, followed by β–α as the most favourable, followed by α–α. β–β is the least favourable interface. A polar fenestration formed at the β–β interface is large enough (more than 5 Å) for hydrated ions to pass through (Fig. 2d and Extended Data Fig. 7) and contributes to the decreased surface

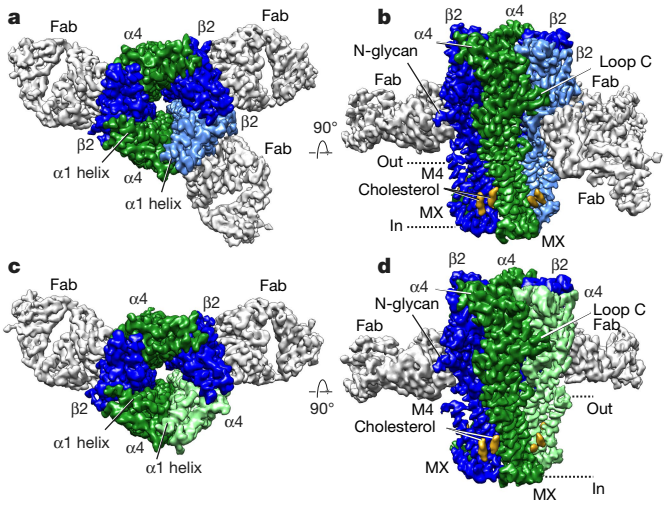
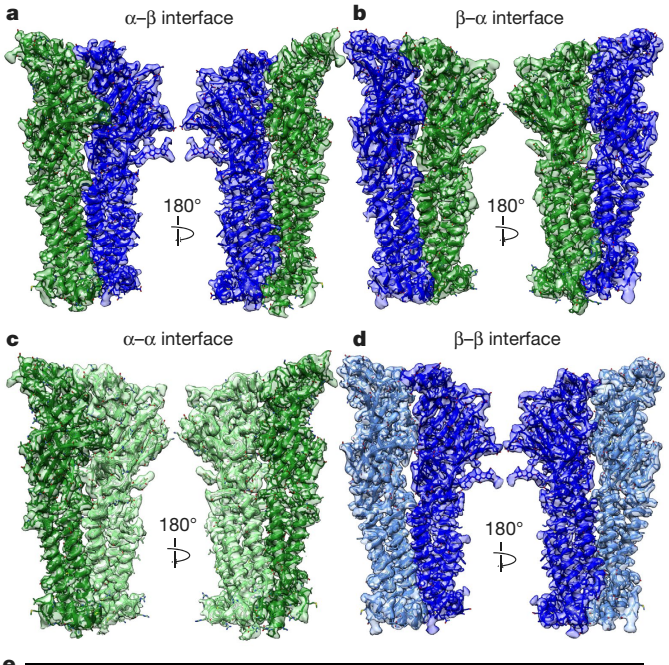


Fig. 1 | Fab-assisted cryo-EM resolves heterogeneity in nicotinic receptor subunit assembly. a, **b**, 2α:3β–Fab complex reconstruction. Fab fragments are light grey, density corresponding to cholesterol is gold, and the α4 and β2 subunits are green and blue, respectively, with a lighter shade of green or blue denoting the subunit position that is unique between the two stoichiometries. **c**, **d**, 3α:2β–Fab complex reconstruction. **b**, **d**, Dashed lines indicate approximate membrane position.

[1]Department of Neuroscience, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [3]Department of Bioengineering and BioX Program, Stanford University, Stanford, CA, USA. [4]Biosciences Division, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. [5]These authors contributed equally: Richard M. Walsh Jr, Soung-Hun Roh. *e-mail: ryan.hibbs@utsouthwestern.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



e

Interface	Interface area (Å^2)	Calculated free energy ΔG (kcal mol^{-1})
2α3β (α–β interface)	2,805	–45
3α2β (α–β interface)	2,797	–44
2α3β (β–α interface)	2,525	–42
3α2β (β–α interface)	2,532	–43
3α2β (α–α interface)	2,923	–40
2α3β (β–β interface)	2,086	–38

Fig. 2 | Interface classes. a–**d**, View from the periphery of the receptor towards the pore axis (left) and from the pore axis out towards the periphery of the receptor (right) for the four types of interface. α4 and β2 subunits are coloured as in Fig. 1. **e**, Calculated interfacial buried surface areas and free energies. For interface classes for which there are two interfaces per assembly (α–β and β–α interfaces), values shown are averages.

area buried at this interface class. These observations suggest that the more favourable α–β and β–α interfaces assemble first, followed by incorporation of the fifth subunit to create the unique interface as the final step in pentamer formation. Superposition of like subunits, both within an assembly and between assemblies, reveals no substantial conformational differences among α-subunits. However, there is a rigid body twist of 4.4° in the ECDs relative to the transmembrane domains (TMDs) of the two β-subunits that compose the unique β–β interface compared to all other β2 subunits (Extended Data Fig. 6d–h). This conformational heterogeneity in β2 subunit structure is likely to add to the instability of the β–β interface. These structural observations suggest that the 3α:2β stoichiometry would be the predominant assembly in the absence of other factors, consistent with observations in rat motor cortex[11] and in expression of recombinant mammalian receptors[9]. In our sample preparation, we exploited nicotine and low temperature, two factors that are known to shift the stoichiometry towards the 2α:3β assembly, during viral transduction to boost overall receptor expression[9].

Analysis of interfaces and subunit conformation suggested an order of subunit incorporation in the observed 2α:3β and 3α:2β assemblies driven by free energy, but did not explain why other theoretically possible subunit stoichiometries were not observed. To address this question, we superimposed the unique interface from each assembly (α–α or β–β) upon itself to build up theoretical homopentameric assemblies, wherein all interfaces except for the final interface would be equivalent to the starting, experimentally observed interface. We then analysed this last, non-experimental interface and found in the 5α assembly a marked increase in surface area buried compared to the experimental α–α interface (3,849 Å^2 versus 2,923 Å^2) and 42 atomic clashes (Extended Data Fig. 8). Put simply, homopentameric α4 assemblies are unfavourable because they pack too tightly. Conversely, in the 5β assembly, at the final, non-experimental interface, we observed a gap between subunits resulting in a large decrease in surface area buried compared to the experimental β–β interface (1,051 Å^2 versus 2,086 Å^2). Thus, β–β interfaces pack such that they create too obtuse an angle to effectively close the pentameric ring. We added one α–β interface into each of these homopentameric assemblies to examine the final theoretically possible subunit stoichiometries and observed a small decrease in the number of clashes in the 4α:1β assembly (39 clashes), and a small increase in the surface area buried at the 1α:4β assembly (1,301 Å^2). We conclude from these comparisons that only two stoichiometries are observed because only one α–α or one β–β interface can be accommodated in the pentamer while maintaining favourable interactions at all interfaces around the pentameric ring.

The ratio of the two observed assemblies of the α4β2 receptor in vivo is tunable and context-specific and has been linked to genetic disorders and addiction[2–5]. Smoking, as well as nicotine added to cultured cells, preferentially upregulates the 2α:3β stoichiometry[2]. Mutations that shift the stoichiometry towards either the 3α:2β or 2α:3β assembly result in autosomal-dominant nocturnal frontal lobe epilepsy[4,5]. A high-penetrance form of this syndrome is linked to a serine-to-leucine mutation in the α4 subunit, at the 10′ position in the M2 pore-lining helix (Extended Data Fig. 6c), which results in a shift in subunit stoichiometry towards the 3α:2β assembly[4]. The mechanism underlying this shift in assembly has been unclear. In the cryo-EM structures, we observe this serine side chain orienting into a hydrophobic pocket in the adjacent subunit, regardless of the identity of that subunit. Mutation from the polar serine to the hydrophobic leucine should be energetically favourable, given the local environment. This substitution would favour assembly of the 3α:2β stoichiometry because with three mutated α-subunits, three interfaces would have improved hydrophobic interactions and fewer unsatisfied hydrogen bonding interactions, contrasted with only two such favourable interfaces in the 2α:3β assembly.

The 3α:2β assembly is classically described as the low-sensitivity isoform of the α4β2 receptor, with half maximal effective concentration (EC_{50}) values for agonist approximately 100-fold higher than for the 2α:3β stoichiometry[9]. We sought to understand this agonist sensitivity difference in the context of receptor structure. The high-sensitivity neurotransmitter sites in the α4β2 receptor are found at the α4–β2 subunit interfaces in ECDs. The α–α interface in the 3α:2β assembly may serve as a low-sensitivity neurotransmitter site, occupation of which is necessary to obtain a full response[12,13]. We included 1 mM nicotine in our preparation to saturate high- and low-sensitivity sites. A comparison of the four classes of interface reveals density for nicotine at the low-sensitivity α–α interface in the 3α:2β stoichiometry in addition to the high-sensitivity α–β interfaces in both stoichiometries of the α4β2 receptor (Fig. 3a, b and Extended Data Fig. 9). All α–β interfaces are qualitatively equivalent and are consistent with the crystal structure[8] in the binding orientation of nicotine and the interactions it makes at this interface class. The binding orientation of nicotine at the α–β neurotransmitter binding sites is distinct from that observed at the α–α binding site. In the α–β interfaces, nicotine is oriented with its long axis approximately parallel to the pore axis, with its pyridine ring oriented away from the membrane. At the α–α interface (Fig. 3a and Extended Data Fig. 9), the pyridine ring of nicotine orients towards the pore axis with the molecule tilted off the membrane normal. This pose of nicotine in the α–α binding site is consistent with crystal structures of other α4β2 agonists bound to a soluble extracellular domain engineered to mimic the α–α binding site in α4β2[14]. This trend suggests that the distinct binding orientation is a meaningful feature of the low-sensitivity α–α binding site. The difference in sensitivity between the classical α–β neurotransmitter binding site (high sensitivity) and the unique α–α binding site

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

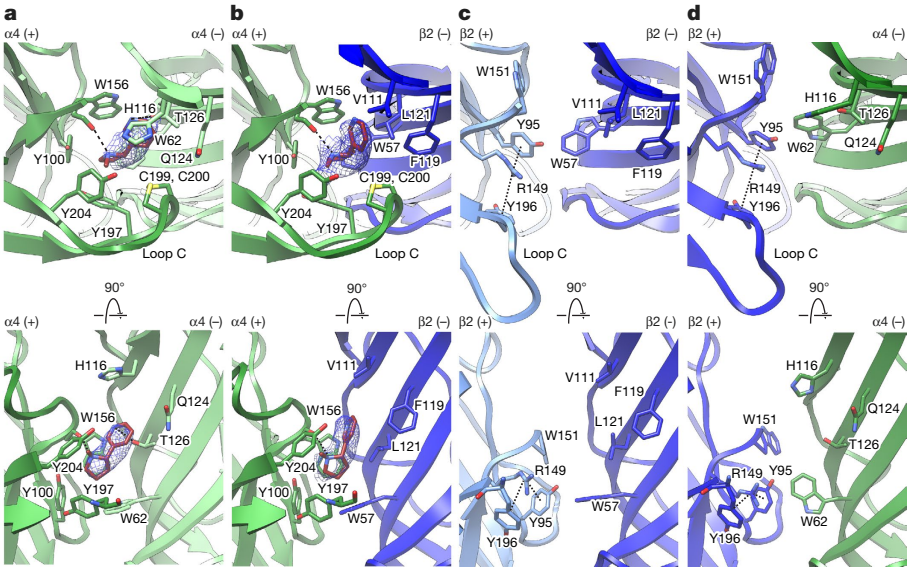


Fig. 3 | Agonist and pseudo-agonist binding sites. a–d, Molecular details of the α–α (**a**) and α–β (**b**) nicotine-binding interfaces and corresponding regions at the β–β (**c**) and β–α (**d**) interfaces. Top row is synaptic perspective viewed parallel to channel axis. Nicotine and interacting residues are shown as sticks. Nicotine density is displayed as blue mesh. In the bottom row, loop C backbone is hidden.

(low sensitivity) has been attributed to three amino acid positions on the complementary (−) face of the binding pocket: V111/F119/L121 for the β2 (−) subunit and H116/Q124/T126 for the α4 (−) subunit[14,15] (Fig. 3c, d and Extended Data Fig. 9). In the α–β site (Fig. 3b and Extended Data Fig. 9), the three hydrophobic residues from the β2 (−) subunit provide a surface of van der Waals contacts that orients the pyridine ring of nicotine, locking it in the aromatic box in a vertical position. In the α–α interface, two of the three variant amino acid positions of α4 (−), T126 and Q124, by the nature of being shorter and polar amino acids, provide a less compact and more hydrophilic site that allows the alternative binding orientation.

Owing in part to limited high-resolution structural information for heteropentameric assemblies, the role of the pseudo-agonist sites, or interface sites in the ECD that do not bind agonist, has been relatively unexplored. Previous crystallographic work revealed the molecular architecture of the pseudo-agonist sites[8], and here we extend this analysis to all interfaces found in the two stoichiometries of the α4β2 nicotinic acetylcholine receptor (Fig. 3c, d). The uniqueness of these pseudo-agonist site interfaces stems from the amino acid sidechains presented by β2 when it occupies the principal (+) side of the subunit interface. Of particular interest is R149, which would electrostatically and sterically impede agonists from binding and is positioned to form cation-π interactions with the sandwiching aromatic residues Y95 and Y196 from β2 (+) (Fig. 3c, d). This intramolecular interaction is conserved among all pseudo-agonist binding interfaces between the two assemblies; furthermore, in the closed, presumably desensitized conformations presented here, all the β–α pseudo-agonist sites in both assemblies appear to be structurally equivalent. We speculate that the guanidium group of R149, which chemically and structurally mimics the nicotinic pharmacophore, may act in part as a covalent agonist, potentiating activation.

The two α4β2 receptor assemblies have different permeation properties: the 3α:2β assembly is about threefold more permeable to Ca^{2+} than the 2α:3β assembly[16] and has a higher single channel conductance[9]. Both permeation pathways are strongly electronegative, as would be expected for a cation-selective channel, with the 3α:2β assembly having a stronger electronegative potential throughout (Fig. 4a, b). These differences in electrostatic potential are localized to the extracellular vestibule and the junction of the ECD and TMD. The stronger electronegative potential seen in the extracellular vestibule of the 3α:2β assembly could not be predicted from sequence alone, as the acidic side chains lining the extracellular vestibule that are thought to affect conductance are conserved between the α4 and β2 subunits[17]. The continuity of the electronegative potential is interrupted in the 2α:3β assembly at the junction between the ECD and TMD, and results from opposing charges at the 20′ position, E268 in 3α:2β and K260 in 2α:3β, in the pore-lining M2 α-helices (Fig. 4c). Point mutations at the 20′ position have demonstrated that having a negative charge at this extracellular entrance to the channel pore is an important determinant of conductance in nicotinic receptors[18] and calcium permeability in cation-selective Cys-loop receptors[16,19]. In the α4β2 receptor specifically, mutational studies with concatemers[16] have shown that increasing the proportion of negative charges at this position correlates directly with increasing permeability to Ca^{2+}. Thus, the difference in charge at the 20′ position is a principal determinant of the different permeation properties of the two receptor subunit assemblies. A caveat to these comparisons is that our electron microscopy construct lacks much of the receptor's intracellular domain, a region that has a role in subunit assembly and ion conductance in the structurally homologous 5-HT_3 receptor[19,20].

Analysis of the transmembrane pore reveals that both α4β2 receptor assemblies taper to a constriction point created by glutamates present in all subunits, located at the −1′ position of the M2 α-helices, where the pore interfaces with the cytosol (Fig. 4c). The 3α:2β assembly has a smaller constriction point, a diameter of 2.4 Å compared to 3.2 Å (Fig. 4d and Extended Data Fig. 10), which is consistent with the more tightly packed α–α interface compared to the β–β interface in the 2α:3β assembly (Fig. 2e and Extended Data Figs. 7, 8). Early studies on permeation through different channel classes suggested that permeation through nicotinic receptors (unlike for K^+ channels) occurs with the ion at least partially hydrated[21]. The smallest physiologically permeant ion is Na^+, with a crystal diameter of about 1.9 Å. Addition of a single equatorial water molecule (diameter 2.8 Å) raises the diameter of the permeant species above the size of both constriction points. With the caveat that the structure was determined in the presence of detergent and is not in a native lipid membrane, the constriction diameter is consistent with the pore representing a non-conducting, agonist-bound desensitized state.

Reconstitution of nicotinic receptors into defined lipid preparations has demonstrated that cholesterol is important for nicotinic receptor function[22]. We tested cholesterol derivatives in thermal stabilization assays[23] and found that the water-soluble cholesterol analogue

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

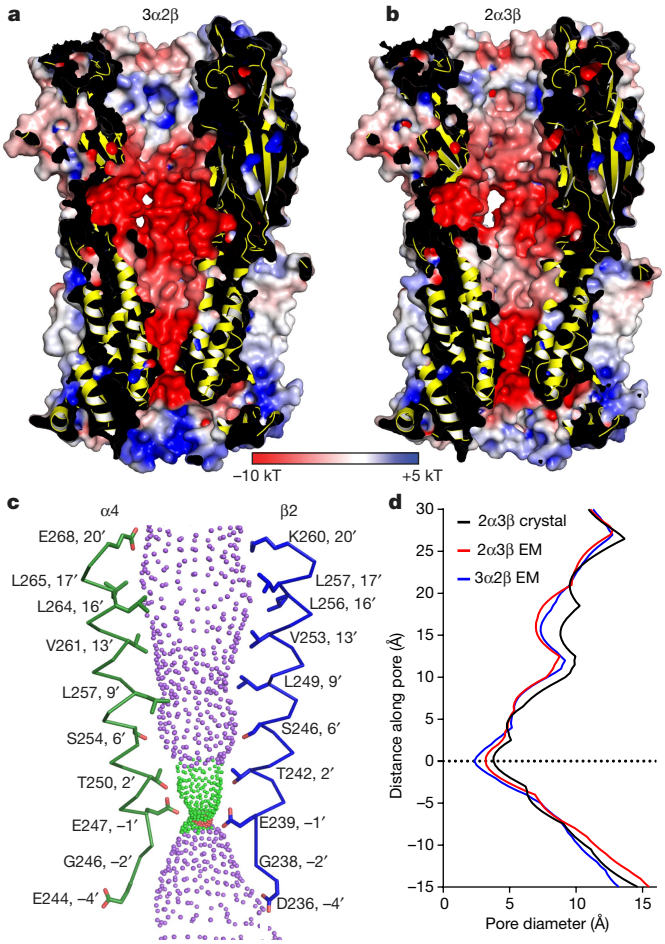


Fig. 4 | Pore and conductance determinants. a, Cutaway of the 3α:2β assembly showing the permeation pathway coloured by electrostatic potential. **b**, As in **a**, but for the 2α:3β assembly. **c**, M2 α-helices from opposing α4 (green) and β2 (blue) subunits from the 3α:2β assembly with sidechains shown for pore-lining residues. Purple spheres indicate pore diameters over 5.6 Å; green are 2.8–5.6 Å; red are less than 2.8 Å. **d**, Pore diameter for the α4β2 receptor 2α:3β and 3α:2β assemblies from the cryo-EM study compared to the crystal structure of the 2α:3β assembly (Protein Data Bank (PDB) accession 5KXI[8]). The zero value along the *y*-axis of the plot is aligned with the α-carbon of the −1′ position. Uncertainty in the position of most acidic sidechains in EM density maps adds ambiguity to the constriction point diameters defined by the −1′ glutamates.

cholesteryl-hemisuccinate (CHS) potently stabilizes the receptor; thus, we included it throughout purification of the receptor. We identified two putative cholesterol molecules per subunit in the electron microscopy density maps (Extended Data Fig. 11). These sausage-shaped densities are found in pairs located at the receptor periphery along the intracellular half of the transmembrane domain, where they flank the subunit interfaces. When viewed at the interface, each putative cholesterol molecule interacts almost exclusively with one subunit and one adjacent cholesterol molecule. Each pair of cholesterol molecules shares a single amino acid contact that bridges these intrasubunit sites (Extended Data Fig. 11c, d). These intrasubunit sites contrast with intersubunit binding sites found for cholesterol and potentiating neurosteroids in $GABA_A$ receptor chimaeras[24,25], and are distinct from the intrasubunit site identified for inhibitory neurosteroids in a $GABA_A$ receptor chimaera[25] (Extended Data Fig. 11f). Comparison of the cholesterol positions in the α4β2 structures suggests that the orientations of the two molecules at an interface are dictated by the identity of the subunit on the principal side. When the principal subunit is β2, the cholesterols are oriented parallel to the pore axis; however, when the principal subunit is α4, the apical ends of both molecules tilt towards the principal face. The cholesterol binding site on the principal subunit is a bowl-shaped surface formed by the junction of the M3, MX and M4 helices. Close examination of this binding site reveals only a single amino acid position that is non-isosteric between the two subunit types, located on the penultimate helical turn in the M3 transmembrane helix. In β2 subunits this residue is a cysteine (C292) and in α4 subunits it is a phenylalanine (F300; Extended Data Fig. 11c–e). The bulky phenylalanine residue is likely to prevent access to the apical portion of the binding region observed when β2 subunits form the principal face. Analysis of human Cys-loop receptor sequences reveals that a small hydrophobic residue that would allow this axial binding mode is conserved in neuronal β2 and β4 subunits as well as the muscle-type β1 subunit. Meanwhile, in all α-subunits from cation-selective receptor types, a bulky hydrophobic residue (F, I, L or M) is conserved at this position. Our assignment of this cholesterol binding site is consistent with mapping studies using photo-activatable cholesterol analogues[26,27].

Here we used single particle cryo-EM coupled with antibody fragments to determine structures of both stoichiometries of the α4β2 nicotinic receptor from one experimental preparation. Comparison of the structures reveals principles underlying subunit assembly and the differences in permeation properties between the two stoichiometries. We observed agonist binding at the non-canonical α–α subunit interface, differentially stabilized by residues unique to this interface class. The structure further allowed mapping of putative cholesterol sites on the surface of the transmembrane domain near the cytosolic junction. This study provides a framework for further biophysical analysis of the allosteric interfacial sites, therapeutic targeting of these sites with modulators akin to benzodiazepines, and more broadly, a general approach for resolving the structural biology of heterogeneous assemblies.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0081-7.

Received: 22 November 2017; Accepted: 23 March 2018;
Published online 2 May 2018.

1. Nemecz, Á., Prevost, M. S., Menny, A. & Corringer, P. J. Emerging molecular mechanisms of signal transduction in pentameric ligand-gated ion channels. *Neuron* **90**, 452–470 (2016).
2. Lester, H. A. et al. Nicotine is a selective pharmacological chaperone of acetylcholine receptor number and stoichiometry. Implications for drug discovery. *AAPS J.* **11**, 167–177 (2009).
3. Moroni, M., Zwart, R., Sher, E., Cassels, B. K. & Bermudez, I. α4β2 nicotinic receptors with high and low acetylcholine sensitivity: pharmacology, stoichiometry, and sensitivity to long-term exposure to nicotine. *Mol. Pharmacol.* **70**, 755–768 (2006).
4. Son, C. D., Moss, F. J., Cohen, B. N. & Lester, H. A. Nicotine normalizes intracellular subunit stoichiometry of nicotinic receptors carrying mutations linked to autosomal dominant nocturnal frontal lobe epilepsy. *Mol. Pharmacol.* **75**, 1137–1148 (2009).
5. Weltzin, M. M., Lindstrom, J. M., Lukas, R. J. & Whiteaker, P. Distinctive effects of nicotinic receptor intracellular-loop mutations associated with nocturnal frontal lobe epilepsy. *Neuropharmacology* **102**, 158–173 (2016).
6. Henderson, R. et al. Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy. *J. Mol. Biol.* **413**, 1028–1046 (2011).
7. Scheres, S. H. Processing of structurally heterogeneous cryo-EM data in RELION. *Methods Enzymol.* **579**, 125–157 (2016).
8. Morales-Perez, C. L., Noviello, C. M. & Hibbs, R. E. X-ray structure of the human α4β2 nicotinic receptor. *Nature* **538**, 411–415 (2016).
9. Nelson, M. E., Kuryatov, A., Choi, C. H., Zhou, Y. & Lindstrom, J. Alternate stoichiometries of α4β2 nicotinic acetylcholine receptors. *Mol. Pharmacol.* **63**, 332–341 (2003).
10. Carbone, A. L., Moroni, M., Groot-Kormelink, P. J. & Bermudez, I. Pentameric concatenated $(\alpha4)_2(\beta2)_3$ and $(\alpha4)_3(\beta2)_2$ nicotinic acetylcholine receptors: subunit arrangement determines functional expression. *Br. J. Pharmacol.* **156**, 970–981 (2009).
11. DeDominicis, K. E. et al. The $(\alpha4)_3(\beta2)_2$ stoichiometry of the nicotinic acetylcholine receptor predominates in the rat motor cortex. *Mol. Pharmacol.* **92**, 327–337 (2017).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

12. Harpsøe, K. et al. Unraveling the high- and low-sensitivity agonist responses of nicotinic acetylcholine receptors. *J. Neurosci.* **31**, 10759–10766 (2011).
13. Mazzaferro, S. et al. Additional acetylcholine (ACh) binding site at α4/α4 interface of $(\alpha4\beta2)_2\alpha4$ nicotinic receptor influences agonist sensitivity. *J. Biol. Chem.* **286**, 31043–31054 (2011).
14. Shahsavar, A. et al. Acetylcholine-binding protein engineered to mimic the α4-α4 binding pocket in α4β2 nicotinic acetylcholine receptors reveals interface specific interactions important for binding and activity. *Mol. Pharmacol.* **88**, 697–707 (2015).
15. Ahring, P. K. et al. Engineered α4β2 nicotinic acetylcholine receptors as models for measuring agonist binding and effect at the orthosteric low-affinity α4-α4 interface. *Neuropharmacology* **92**, 135–145 (2015).
16. Tapia, L., Kuryatov, A. & Lindstrom, J. Ca^{2+} permeability of the $(\alpha4)_3(\beta_2)_2$ stoichiometry greatly exceeds that of $(\alpha4)_2(\beta2)_3$ human acetylcholine receptors. *Mol. Pharmacol.* **71**, 769–776 (2007).
17. Hansen, S. B., Wang, H. L., Taylor, P. & Sine, S. M. An ion selectivity filter in the extracellular domain of Cys-loop receptors reveals determinants for ion conductance. *J. Biol. Chem.* **283**, 36066–36070 (2008).
18. Imoto, K. et al. Rings of negatively charged amino acids determine the acetylcholine receptor channel conductance. *Nature* **335**, 645–648 (1988).
19. Livesey, M. R. et al. Structural determinants of Ca^{2+} permeability and conduction in the human 5-hydroxytryptamine type 3A receptor. *J. Biol. Chem.* **283**, 19301–19313 (2008).
20. Kelley, S. P., Dunlop, J. I., Kirkness, E. F., Lambert, J. J. & Peters, J. A. A cytoplasmic region determines single-channel conductance in 5-HT3 receptors. *Nature* **424**, 321–324 (2003).
21. Dwyer, T. M., Adams, D. J. & Hille, B. The permeability of the endplate channel to organic cations in frog muscle. *J. Gen. Physiol.* **75**, 469–492 (1980).
22. Baenziger, J. E., Domville, J. A. & Therien, J. P. D. The role of cholesterol in the activation of nicotinic acetylcholine receptors. *Curr. Top. Membr.* **80**, 95–137 (2017).
23. Hattori, M., Hibbs, R. E. & Gouaux, E. A fluorescence-detection size-exclusion chromatography-based thermostability assay for membrane protein precrystallization screening. *Structure* **20**, 1293–1299 (2012).
24. Miller, P. S. et al. Structural basis for $GABA_A$ receptor potentiation by neurosteroids. *Nat. Struct. Mol. Biol.* **24**, 986–992 (2017).
25. Laverty, D. et al. Crystal structures of a $GABA_A$-receptor chimera reveal new endogenous neurosteroid-binding sites. *Nat. Struct. Mol. Biol.* **24**, 977–985 (2017).
26. Hamouda, A. K., Chiara, D. C., Sauls, D., Cohen, J. B. & Blanton, M. P. Cholesterol interacts with transmembrane α-helices M1, M3, and M4 of the *Torpedo* nicotinic acetylcholine receptor: photolabeling studies using $[^3H]$Azicholesterol. *Biochemistry* **45**, 976–986 (2006).
27. Corbin, J., Wang, H. H. & Blanton, M. P. Identifying the cholesterol binding domain in the nicotinic acetylcholine receptor with $[^{125}I]$azido-cholesterol. *Biochim. Biophys. Acta* **1414**, 65–74 (1998).

Acknowledgements We thank D. Cawley at OHSU for production of monoclonal antibodies, X. Bai for EM discussion, Y. Jiang for the use of the oocyte rig, C. Noviello for assistance in EM data collection and all members of the Hibbs Laboratory and M. Horvath for discussion. Cryo-EM data were collected at the UT Southwestern Medical Center Cryo-Electron Microscopy Facility, which is funded in part by CPRIT Core Facility Support Award RP170644. We thank D. Nicastro and Z. Chen for support in facility access and data acquisition and W. Chiu for cryo-EM training and resources in the National Center for Macromolecular Imaging (NCMI) at Baylor College of Medicine. NCMI is supported by NIH Grants P41GM103832 and R01GM079429. R.M.W. acknowledges support from the Sara and Frank McKnight Fund for Biochemical Research and the NIH (T32GM008203). R.E.H. is supported by a McKnight Scholar Award, The Welch Foundation (I-1812) and the NIH (DA037492, DA042072, and NS095899).

Reviewer information *Nature* thanks R. Aricescu, R. Dutzler and M. Jansen for their contribution to the peer review of this work.

Author contributions R.M.W. expressed and purified the protein, made the EM samples, collected the EM data, processed the EM data with assistance from S.-H.R., built and refined the models and performed the binding assays. A.G. processed an initial EM dataset revealing the presence of two stoichiometries. C.L.M.-P. performed the mAb and Fab characterization. J.T. performed the electrophysiology. R.M.W. and R.E.H. wrote the manuscript with assistance from S.-H.R.; R.E.H. oversaw all aspects of the project. All authors gave feedback on the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0081-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0081-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.E.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Nicotinic receptor expression and purification. The human α4 and β2 nicotinic receptor genes were provided by J. Lindstrom at the University of Pennsylvania. Deletion constructs used for expression and structural analysis lack a large portion of the intracellular domain of the receptor, as previously described[8]. These genes encode the native signal peptides, residues 1–338 and 556–601 in the α4 subunit and residues 1–330 and 417–477 in the β2 subunit (residue numbering here is for the wild-type mature, signal-peptide-cleaved protein sequence). To promote good biochemical behaviour a Glu-Arg linker was inserted in the MX–M4 junction, between Phe559 and Ser560 in the α4 subunit and between Gln420 and Ser421 in the β2 subunit. For purification purposes a Strep-tag was inserted at the C terminus of the β2 subunit preceded by a Ser-Ala linker. The genes were subcloned into the pEZT bacmam expression vector[28] and viral titration and protein expression in GnTI- HEK cells (ATCC CRL-3022, not mycoplasma tested) were performed as previously described[28] with modifications. We transduced 3.2 l of suspension GnTI- cells with multiplicities of infection (MOIs) of 0.5:0.5 for the α4 and β2 subunits, respectively, to promote expression of two receptor assemblies. Nicotine (Sigma-Aldrich) and sodium butyrate (Sigma-Aldrich) were added at the time of transduction to 0.1 mM and 3 mM, respectively. At the time of transduction, suspension cells were moved to 30 °C and 8% CO_2. Cells were collected after ~72 h by centrifugation, resuspended in 20 mM Tris, pH 7.4, 150 mM NaCl (TBS buffer), 1 mM nicotine and 1 mM phenylmethanesulfonyl fluoride (Sigma-Aldrich), and disrupted using an Avestin Emulsiflex. Lysed cells were centrifuged for 15 min at 10,000*g*; supernatants containing membranes were centrifuged for 2 h at 186,000*g*. Membrane pellets were mechanically homogenized and solubilized for 1 h at 4 °C in a solution containing TBS, 40 mM *n*-dodecyl-β-D-maltopyranoside (DDM; Anatrace), 1 mM nicotine and 0.2 mM CHS (Tris salt, Anatrace). Solubilized membranes were centrifuged for 40 min at 186,000*g* then passed over high capacity Strep-Tactin (IBA) affinity resin. The resin was washed with Size Exclusion Chromatography (SEC) buffer containing TBS, 1 mM DDM, 1 mM nicotine, 0.2 mM CHS and 1 mM TCEP (Thermo Fisher Scientific) and eluted in the same buffer containing 5 mM desthiobiotin (Sigma-Aldrich).

Generation of monoclonal antibodies and Fab fragments. The 12H2 monoclonal antibody (mAb) against the α4β2 nicotinic receptor (IgG1, κ) was raised by D. Cawley (Monoclonal Core, Vaccine and Gene Therapy Institute, OHSU) using standard methods. Hybridomas were generated by immunization of mice (female BALB/c) with α4β2 receptor protein purified in DDM, followed by fusion of splenocytes with mouse myeloma cells and selection of clones by enzyme-linked immunosorbent assay. The mAb was purified from hybridoma cell supernatant by cation exchange and protein A affinity chromatography. High affinity and specificity of the antibody for properly folded pentameric α4β2 receptor was assessed using Fluorescence-SEC (FSEC) with GFP-tagged receptor (shift in elution volume) and by western blot (no binding). The top candidate antibodies were then more thoroughly assessed using FSEC monitoring tryptophan fluorescence with purified α4β2 receptor and Fab fragments to ensure fidelity of the α4β2–Fab complex and to estimate receptor:Fab stoichiometry. α4β2–Fab complexes were co-purified, then re-injected in analytical FSEC to test for dissociation under these dilute conditions over a Sepax SRT SEC-500 (4.6 × 300 mm) gel filtration column. Fabs that did not dissociate were qualitatively deemed to have sufficiently high affinity to use in EM studies. The 12H2 Fab fragment used here was generated by papain cleavage of whole antibody at a final concentration of 1 mg/ml for 2 h at 37 °C in 50 mM $NaPO_4$, pH 7.0, 1 mM EDTA, 10 mM cysteine and 1:100 (w:w) papain. Digestion was quenched using 30 mM iodoacetamide at 25 °C for 10 min. Fab was purified by cation exchange using a Hi-Trap SP column in 10 mM sodium citrate pH 5.0 and a NaCl gradient elution. Cloning and sequencing of Fab heavy and light chain genes were performed from mouse hybridoma cells.

Cryo-EM sample preparation. Purified α4β2 from affinity chromatography was mixed with Fab in a 1:2 (w:w) ratio and injected over a Superose 6 Increase 10/300 GL column (GE) equilibrated in SEC buffer. Peak fractions were assayed by SEC, monitoring tryptophan fluorescence. The peak fraction was concentrated fourfold to an A280 of 5.9 in SEC buffer, ultracentrifuged, and used directly to prepare EM grids. Three microlitres of purified α4β2-nicotine-Fab complex was applied to glow-discharged gold R1.2/1.3 300 mesh holey carbon grids (Quantifoil) and immediately blotted for 4 s at ~100% humidity and 4 °C, then plunge-frozen into liquid ethane cooled by liquid nitrogen using a Vitrobot Mark IV (FEI).

Cryo-EM image collection and processing. Two separate data sessions were taken on a Titan Krios electron microscope (FEI) operated at 300 kV. Images were recorded on a K2 Summit direct electron detector (Gatan) equipped with GIF quantum energy filter (20 e^-V) (Gatan) in super-resolution mode (super-resolution pixel size: 0.535 Å per pixel). In the first session, 3,323 images were collected and each micrograph in the first session was exposed for 15 s with a dose rate of ~5 $e^-/A^2/s$ (total specimen dose, ~75 e^-/A^2), and 30 frames were captured per micrograph. In the second session, 2,588 images were collected and each micrograph in the second session was exposed for 10 s with ~5 $e^-/A^2/s$ dose rate (total specimen dose, ~50 e^-/A^2), and 40 frames were captured per micrograph. Images were recorded using the automated-acquisition program EPU (FEI) with defocus values varied from 0.5 μm to 4.5 μm.

Dose-fractionated images (movies) were gain normalized, 2 × Fourier binned (resulting in a pixel size of 1.07 Å per pixel), aligned, dose-weighted and summed using Unblur[29]. Defocus values were estimated using GCTF[30]. A total of 5,166 micrographs were manually selected and used for further analysis (2,836 out of 3,323 from the first session and 2,330 of 2,588 from the second session). Approximately 100 particles were manually picked and subjected to reference-free 2D class averaging in RELION 2.1[31,32]. Resulting references were then used for auto-picking in RELION 2.1 for a subset of 10 images, then the classification process was repeated to obtain good references for auto-picking on all images. Following auto-picking, micrographs were manually inspected, and false-positive particles were manually removed. Curation of auto-picked particles resulted in a total of 649,773 particles. 516,450 particles were selected following 2D classification in RELION. Two-hundred particles from 30 different 2D classes (6,000 particles in total) were used to generate an ab initio model in RELION. 3D classification using eight classes resulted in six good classes (403,246 particles) that could be divided into to two distinct categories based on Fab occupancy and subunit-defining features, representing the 3α:2β (139,551), and 2α:3β (263,695) stoichiometries of the α4β2 receptor. 3D refinement of these two pooled particle sets yielded reconstructions of the overall receptor-Fab complexes at ~4.0 Å resolution. The Fab molecules, in particular the peripheral constant domains, were poorly resolved, and of less biological interest to us, so we focused our effort on the receptor. Removing the Fab regions from the global resolution calculations results in reconstructions for the 3α:2β assembly at 3.7 Å and for the 2α:3β assembly at 3.4 Å resolution, consistent with side chain features throughout the density maps. All global resolutions were estimated by applying a soft mask around the protein density and the gold-standard Fourier shell correction (FSC) = 0.143 criterion. ResMap[33] was used to estimate the local resolution.

Model building, refinement and validation. The higher resolution 2α:3β assembly was built first. The X-ray coordinates (PDB accession code 5KXI) were docked into the density map. Secondary structural features that are distinct between the two stoichiometries (loop C and α1 helix, Fig. 1) were used to guide the initial rigid body fit using UCSF Chimera[34]. SwissModel[35] was used to generate a homology model of the Fab light chain using PDB entry 4TPR and for the heavy chain using PDB entry 2ZJS. These chains were docked into the EMD density at one Fab site. Manual adjustments of the receptor-Fab structure were then performed in Coot[36]. The ECD and transmembrane domain (TMD) halves of each subunit, and the variable and constant halves of each Fab, were rigid body fitted into the density map. The epitope binding loops of the Fabs were rebuilt into unambiguous density at the site of interaction with the receptor, and then this first Fab molecule was copied into the two additional sites in the 2α:3β assembly. The map quality for the receptor exceeded that from the earlier X-ray structure, which allowed for more residues to be built at the N termini and in the linkers connecting the MX amphipathic helix with the M4 helix. There was unclear density and thus a gap in the connectivity in the MX-M4 linker for 7 residues in α4 subunits and 10 residues in β2 subunits. There was improved density for side chains and better connectivity in the ECD loops, in particular loop C in the β-subunits. After rebuilding in Coot, global real space coordinate and B factor refinement were performed in Phenix[37].

The cryo-EM density maps drew our attention to a notable conformational difference in the loop C of β2 subunits compared to that modelled in the crystallographic structure[8]. We observed continuous density for this substructure in all β2 subunits in EM maps that resolve loop C in a conformation bending out and away from the subunit interface (Fig. 3 c, d and Extended Data Fig. 6a–c). This conformation was independent of whether Fab was bound to the adjacent (−) subunit or not (Extended Data Fig. 1e, f). In the crystal structure, loop C in β2 subunits was modelled in a conformation packed against the (−) subunit, roughly akin to the conformation of loop C in α4 subunits. The crystallographic modelling was based on clear electron density for loop C in one β2 subunit that was at a crystal contact (Extended Data Fig. 6a–c). We thus suggest that an open conformation of loop C in β2 subunits is the proper one.

The validation to test for overfitting of the model was performed as previously described[38]. In brief, the atom positions of the final refined model were randomly displaced by a maximum of 0.5 Å using PDBSET in the CCP4 suite[39]. This perturbed model was then refined in Phenix in real space against the first half map of the reconstruction comprising 50% of the particles. A map versus model FSC comparison was made for this model versus the map used in its refinement ('work'), as well as the same model versus the half map not used in refinement ('free'). The FSC curves of map versus model agree well (Extended Data Figs. 4, 5).

Sequences used in alignments were retrieved from the UniProtKB database[40]. Sequence alignments were made using PROMALS3D[41]. Subunit interfaces includ-

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ing free energies were analysed using the PDBePISA server[42]. Pore diameters were calculated using HOLE[43]. Domain motions were analysed using DynDom[44]. Ligand interactions were characterized by visual inspection and analysis using LIGPLOT[45]. Structural figures were made using UCSF-Chimera[34] and PyMOL (Schrodinger, LLC) including the APBS electrostatics plugin[46]. Structural biology software packages were compiled by SBGrid[47].

Density maps were displayed in Chimera at a threshold of 0.0306 in Figs. 1, 2. In Fig. 3, ligand density is displayed in Chimera at a threshold of 0.0403 for the α–α interface, and 0.05 for the α–β interface.

Radioligand binding. Experiments to test the effect of Fab on [^{3}H]-nicotine (PerkinElmer, 79.8 Ci/mmol) binding to the α4β2 receptor were performed with protein purified as for EM (that is, a mixture of stoichiometries) but in the absence of Fab and nicotine. The concentration of binding sites was kept at 2.5 nM after a preliminary experiment to determine optimal receptor concentration. For the + Fab samples, Fab was added in excess (5 μM). In addition to the protein, the binding assay conditions included 20 mM Tris pH 7.4, 150 mM NaCl, 1 mM DDM, and 1 mg/ml streptavidin-YiSi scintillation proximity assay beads (SPA; GE Healthcare Life Sciences). Non-specific signal was determined in the presence of 100 μM [^{1}H]-nicotine; all data shown are from background-subtracted measurements. Experiments were performed in triplicate. All data were analysed using Prism 6 software (GraphPad) with variable Hill slope.

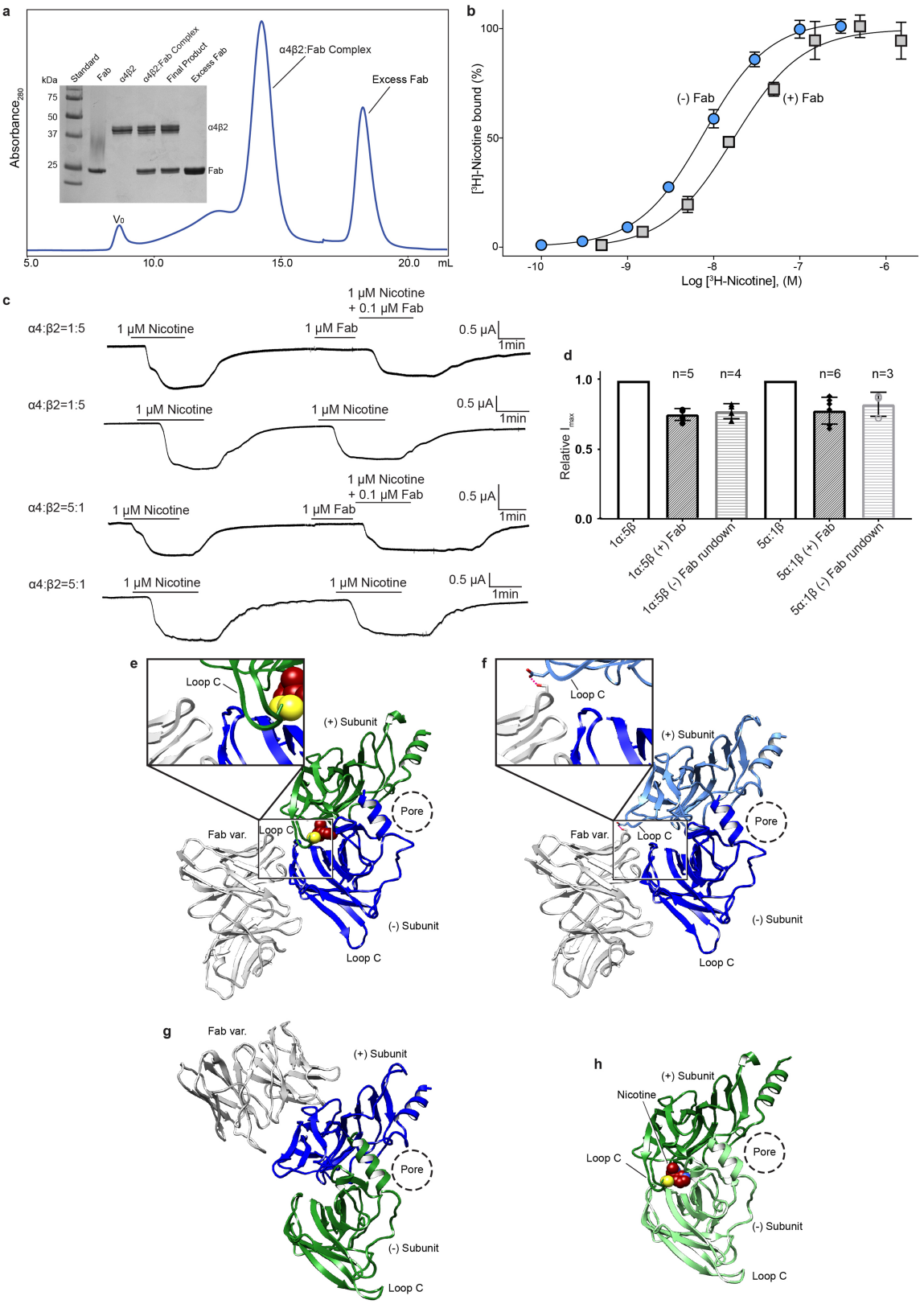
Electrophysiology. The open reading frames of α4 and β2 EM constructs described above were subcloned into the pGH19 vector[48]. cRNAs of α4 and β2 constructs were synthesized in vitro using the MEGAscript T7 Transcription Kit (Thermo Fisher). *Xenopus laevis* ovaries purchased from NASCO (LM00935MX) were treated with 1–2 mg/ml collagenase I (Gibco) in Barth's solution (in mM, 88 NaCl, 1 KCl, 2.4 $NaHCO_3$, 0.82 $MgSO_4$, 0.33 $Ca(NO_3)_2$, 0.68 $CaCl_2$, 10 Hepes, pH 7.4) for 1–1.5 h, and washed with ND96 solution (in mM, 96 NaCl, 2 KCl, 1 $MgCl_2$. 1.8 $CaCl_2$, and 5 Hepes pH 7.4). Stages V–VI oocytes were collected and kept at 17 °C in ND96 solution with 25 μg/ml ampicillin and 0.1 mg/ml gentamycin. Oocytes were injected with 1–10 ng cRNA mixture in the ratio 1:5 or 5:1 of α4:β2. Two to three days after injection of cRNA, receptor activity was measured by two-electrode voltage clamp recording in ND96 solution using Oocyte Clamp OC-725 amplifier (Warner Instruments), and the signal was digitized with a Digidata 1440 A (Molecular Devices). Data were analysed using Clamp 10.3. Pipettes were filled with 1 M KCl and showed resistances of 0.8–2.0 MΩ. Ligands were dissolved in ND96.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Atomic coordinates of the two α4β2 receptor–nicotine–Fab complexes have been deposited in the Protein Data Bank (PDB; 2α:3β, 6CNJ; 3α:2β, 6CNK). The cryo-EM density maps have been deposited in the Electron Microscopy Data Bank (2α:3β, 7535; 3α:2β, 7536).

28. Morales-Perez, C. L., Noviello, C. M. & Hibbs, R. E. Manipulation of subunit stoichiometry in heteromeric membrane proteins. *Structure* **24**, 797–805 (2016).
29. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4**, e06980 (2015).
30. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
31. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. D* **73**, 496–502 (2017).
32. Scheres, S. H. A Bayesian view on cryo-EM structure determination. *J. Mol. Biol.* **415**, 406–418 (2012).
33. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
34. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
35. Biasini, M. et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**, W252–W258, (2014).
36. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
37. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
38. Amunts, A. et al. Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).
39. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
40. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
41. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).
42. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
43. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360, 376 (1996).
44. Hayward, S. & Lee, R. A. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *J. Mol. Graph. Model.* **21**, 181–183 (2002).
45. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8**, 127–134 (1995).
46. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98**, 10037–10041 (2001).
47. Morin, A. et al. Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).
48. Venkatachalan, S. P. et al. Optimized expression vector for ion channel studies in *Xenopus* oocytes and mammalian cells using alfalfa mosaic virus. *Pflugers Arch.* **454**, 155–163 (2007).
49. Galzi, J. L., Bertrand, S., Corringer, P. J., Changeux, J. P. & Bertrand, D. Identification of calcium binding sites that regulate potentiation of a neuronal nicotinic acetylcholine receptor. *EMBO J.* **15**, 5824–5832 (1996).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

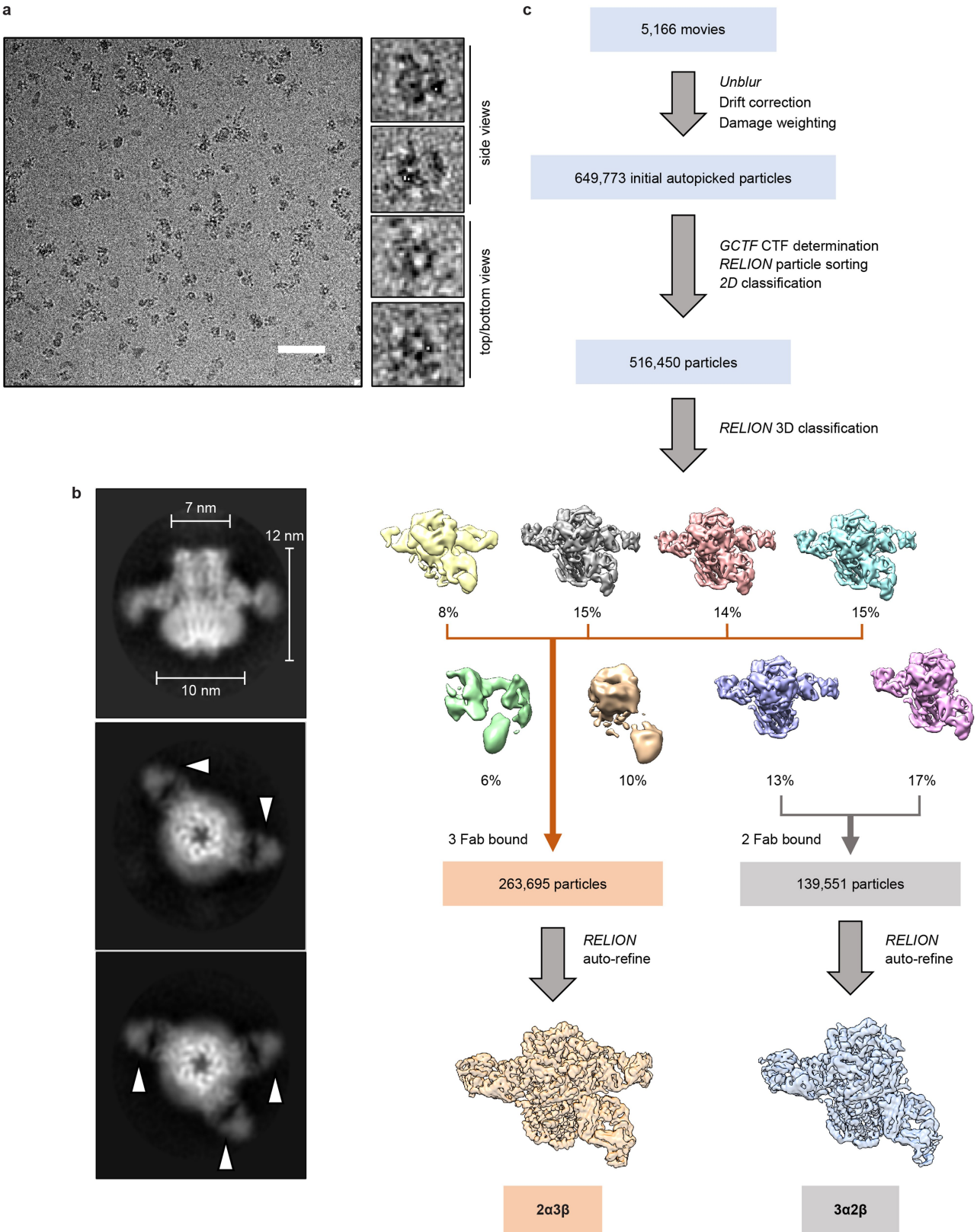


Extended Data Fig. 1 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

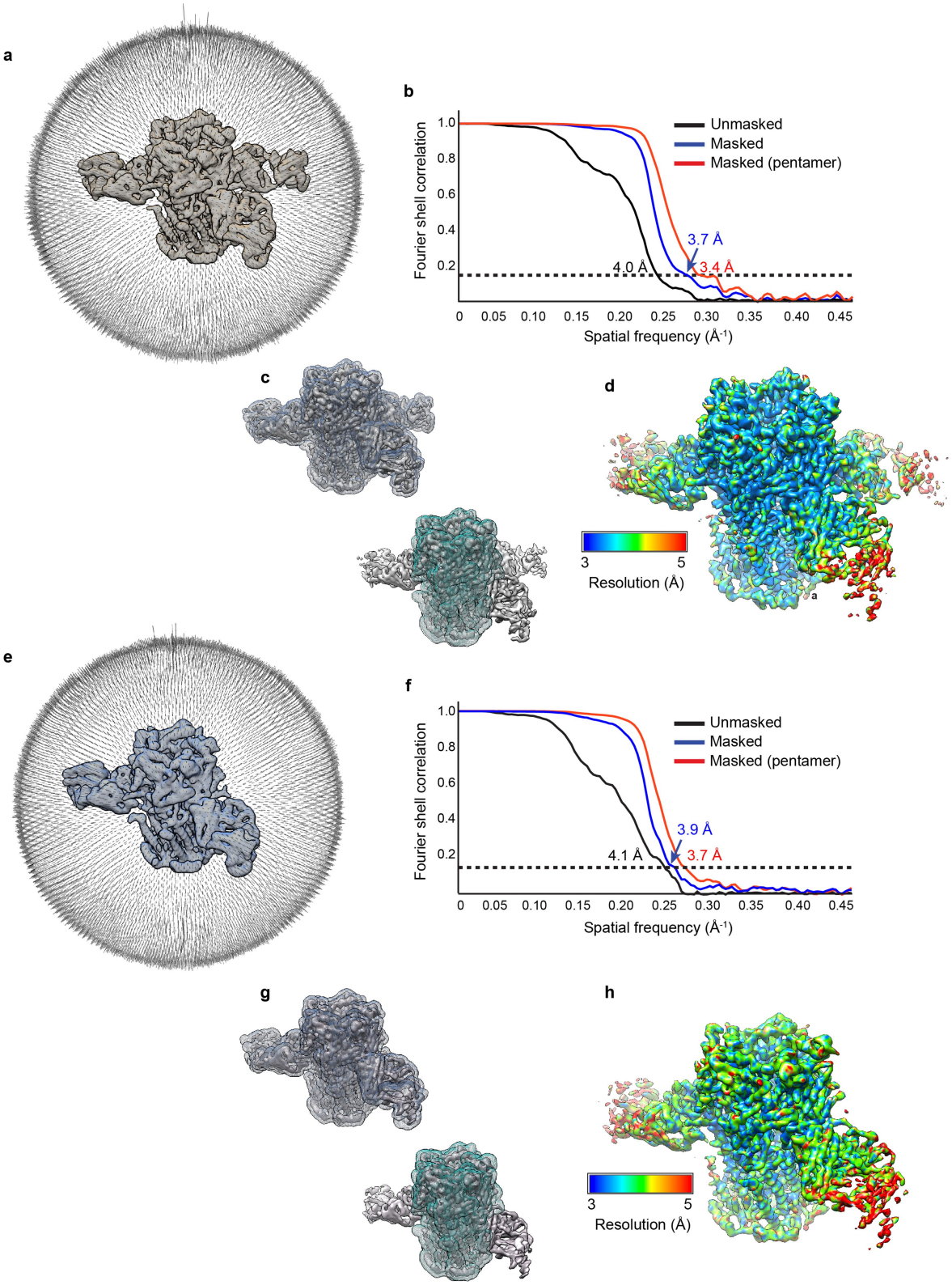
Extended Data Fig. 1 | Biochemistry, binding, electrophysiology and α4β2–Fab interactions. a, Size-exclusion chromatogram of α4β2–Fab complex and SDS–PAGE analysis of complex purification (V_0, void volume). **b**, Saturation binding experiments with [^{3}H]-nicotine and a mixture of receptor subunit stoichiometries. Grey squares denote samples with receptor plus Fab and blue circles denote samples with receptor alone. Receptor alone $K_d = 7.7$ nM (95% confidence interval (CI), 6.9–8.6 nM) and receptor plus Fab $K_d = 17.0$ nM (95% CI, 10.2–26.6 nM). Receptor alone and receptor plus Fab both exhibited a Hill slope of approximately 1 (1.14 and 1.09, respectively). Plotted results are from a representative experiment performed in triplicate. **c**, Representative two-electrode voltage clamp (−60 mV) recordings of oocytes injected with cRNAs for the α4 and β2 subunits in ratios to bias assembly,1α:5β (top two traces) and 5α:1β (bottom two traces) for the 2α:3β and 3α:2β assemblies, respectively. The experiments were performed in the presence and absence of Fab to assess the effect of Fab on receptor gating. Each oocyte was perfused with 1 μM nicotine solution for about 2 min, washed with bath solution for 10 min and then perfused for about 2 min with 1 μM nicotine ± Fab. For samples where Fab was included, 1 μM final [Fab] was added directly to the oocyte bath following the first perfusion and allowed to incubate for 1.5 min before perfusing with a second solution containing 1 μM nicotine + 0.1 μM Fab. **d**, Bar graph quantifying peak currents before and after adding Fab. Currents were normalized to the amplitude of the first nicotine application. Appreciable current rundown was observed (assessed by applying nicotine without Fab as the second application). Change in peak current is similar in the presence and absence of Fab for both stoichiometries, suggesting no substantial effect on gating. *n*, number of oocytes; error bars, s.d. from mean. **e**–**h**, Fab–β-subunit interactions at the α–β, β–β, α–β, and α–α interfaces, respectively. With one exception, the Fab molecules interact exclusively with a single β-subunit. The exception is at the β–β interface, where the Fab on the complementary β-subunit forms one potential interaction with the preceding β-subunit. This interaction is displayed as a dashed line (**f**, inset). The conformation of the principal loop C at the β–β interface is indistinguishable from those where Fab is not interacting, suggesting that Fab does not affect the loop C conformation. Subunits are coloured as in Fig. 1

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



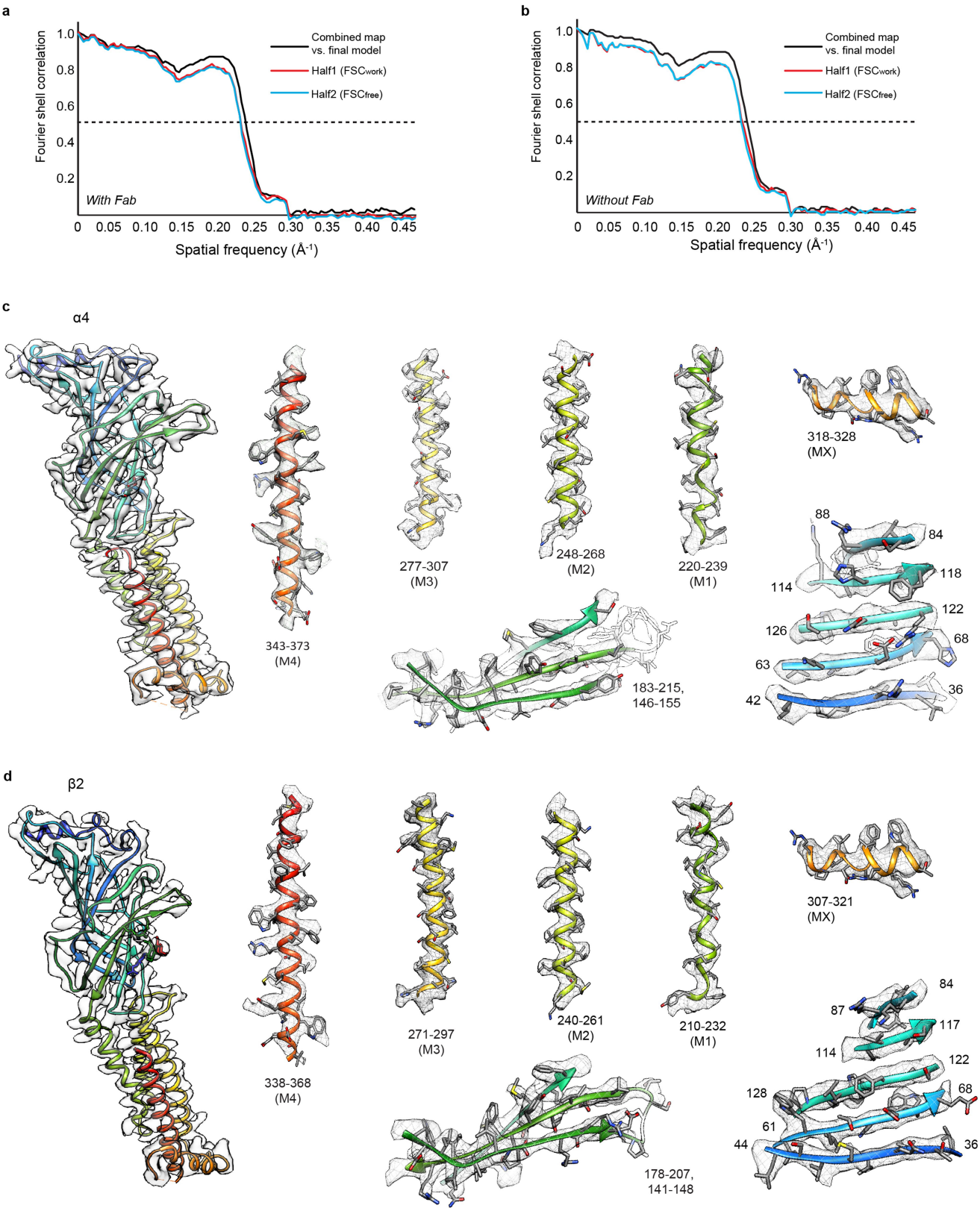
Extended Data Fig. 2 | Cryo-EM image processing procedure. **a**, Representative micrograph of the α4β2–Fab complex (scale bar, 50 nm). Boxed regions on the right are magnified to show representative particle images (top/bottom and side views). **b**, Images of selected 2D classes from reference-free 2D classification by RELION. White arrowheads, Fab fragments. **c**, Overview of the image processing procedure (see Methods).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



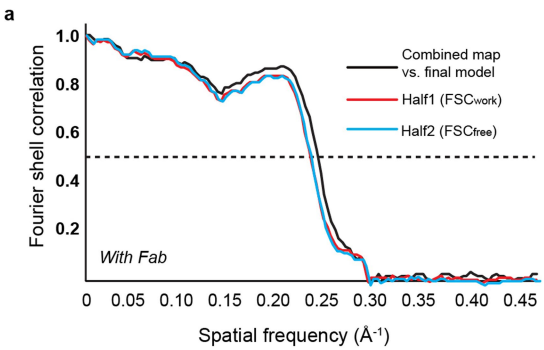
Extended Data Fig. 3 | Three-dimensional reconstructions of the 2α:3β and 3α:2β assemblies. a, Angular distribution histogram of 2α:3β assembly projections. **b**, FSC of 2α:3β assembly maps before (black) and after (blue and red) masking. Two soft masks were used: one (red) mask that included only the receptor and one (blue) mask that included the whole 2α:3β–Fab complex. When a mask was used, the FSC curve was corrected for masking effects during the RELION post-processing procedure. Masks used in FSC calculations are shown in **c** for the whole 2α:3β–Fab complex (top) and receptor only (bottom), superimposed on the map in question. **d**, Local resolution of the 2α3β–Fab reconstruction estimated by ResMap[33]. Shown is the combined map, which has not been sharpened or filtered. **e**–**h**, As in **a**–**d** but for the 3α:2β assembly.

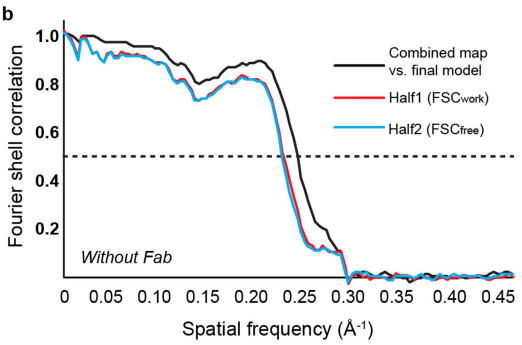
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

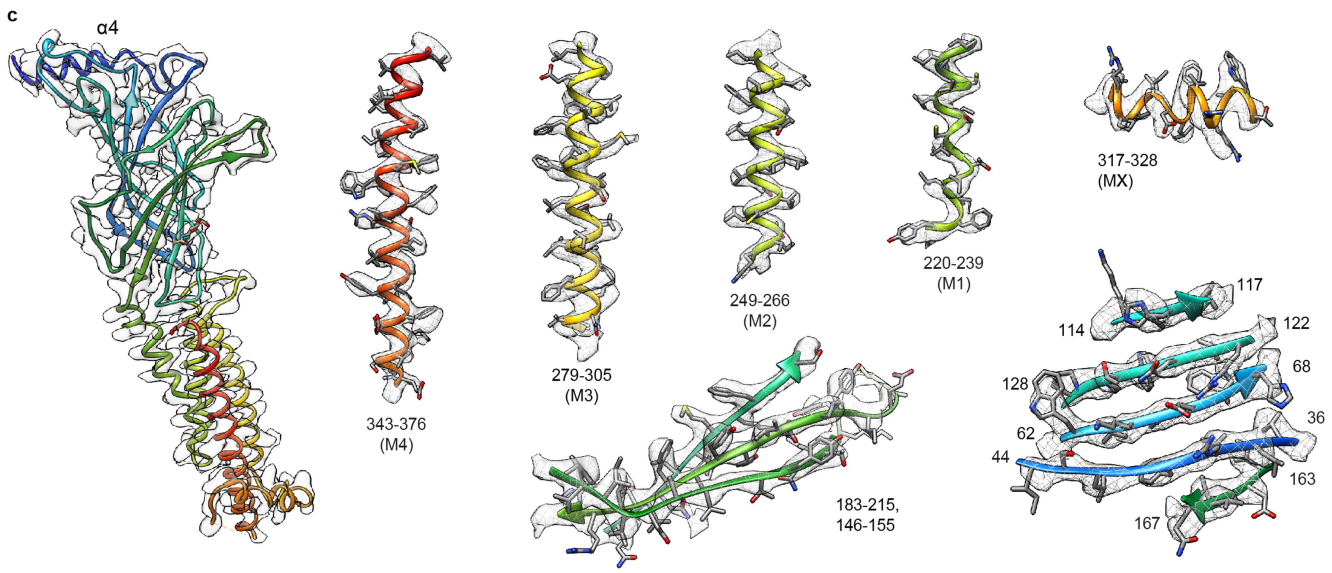


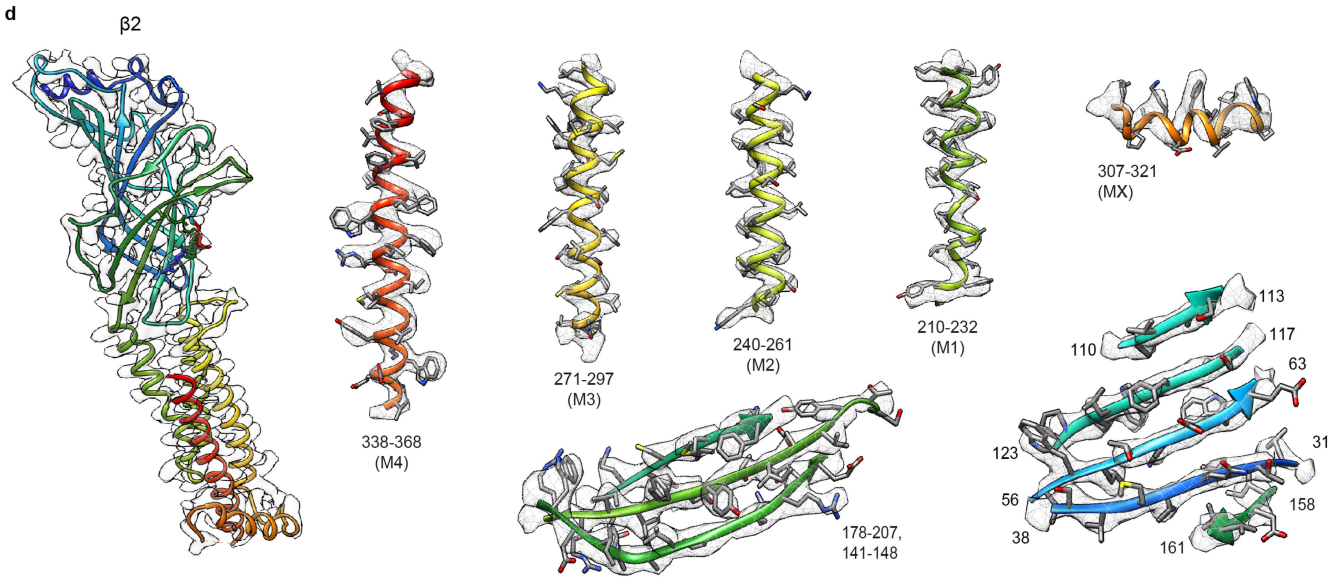
Extended Data Fig. 4 | 3α:2β assembly model-map validation. a, **b**, FSC curves for cross-validation between the maps and the models with (**a**) and without Fab fragments (**b**) for the 3α:2β assembly. Curves for final model versus summed map (full) in black, for model versus half map (work) in red, and for model versus half map not used for refinement (free) in blue. For validation of receptor alone (without Fab), maps were segmented to exclude the Fab fragments. **c**, EM density segments of the 3α:2β assembly for a representative α4 subunit. Density map and model for α4 subunit (left). Representative density for extracellular and transmembrane spanning regions (right). Regions are numbered and helices are labelled. Maps were sharpened with a single *B* factor at −150 Å^2. **d**, As in **c**, but for a representative β2 subunit.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



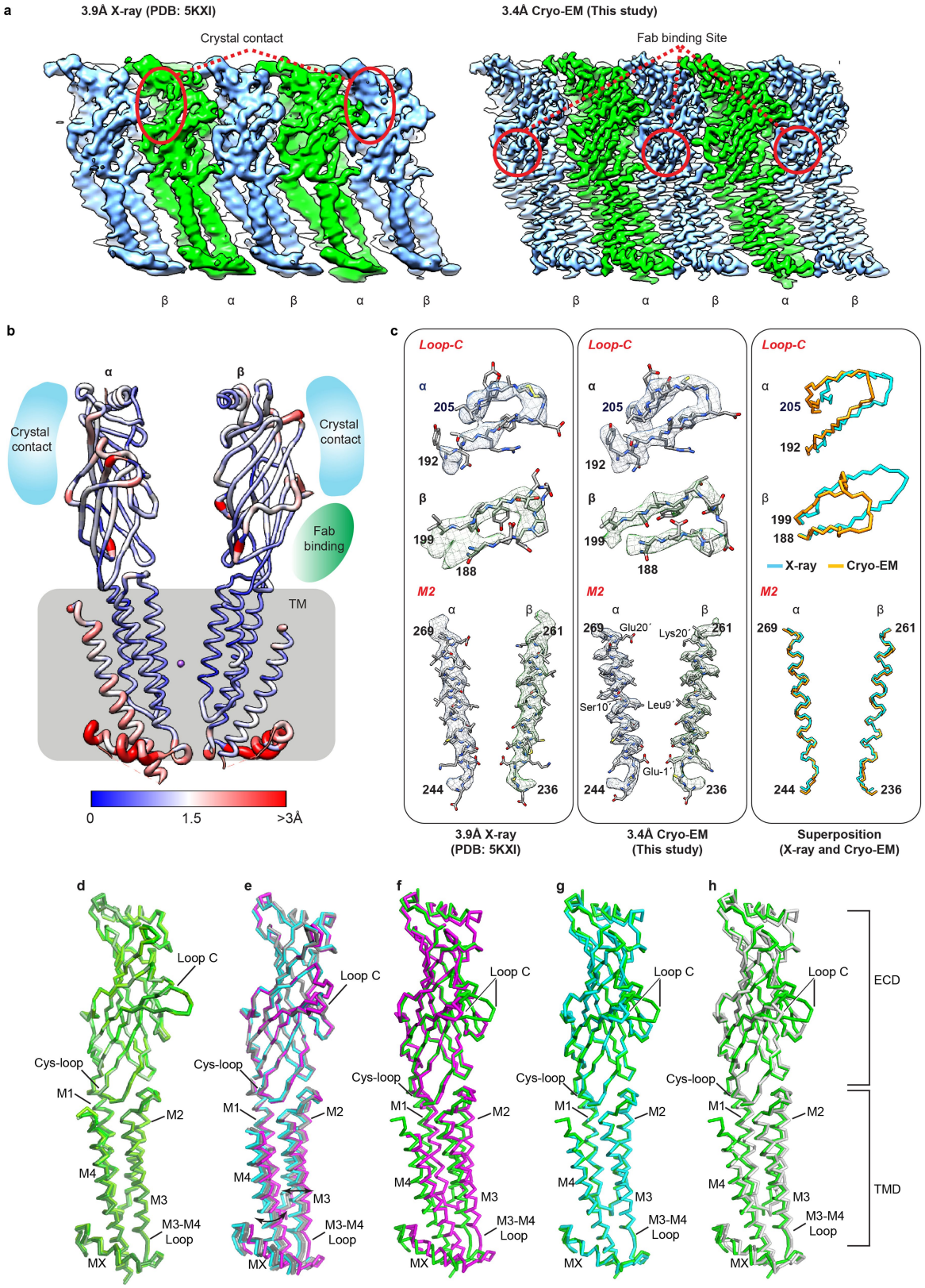






Extended Data Fig. 5 | 2α:3β assembly model–map validation. a, **b**, FSC curves for cross-validation between the maps and the models with (**a**) and without Fab fragments (**b**) for the 2α:3β assembly. Curves for final model versus summed map (full) in black, for model versus half map (work) in red, and for model versus half map not used for refinement (free) in blue. For validation of receptor alone (without Fab) maps segmented to exclude the Fab fragments were used. **c**, EM density segments of the 2α:3β assembly for a representative α4 subunit. Density map and model for α4 subunit (left). Representative density for extracellular and transmembrane spanning regions (right). Regions are numbered and helices are labelled. Maps were sharpened with a single *B* factor at −170 Å^2. **d**, As in **c**, but for a representative β2 subunit.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

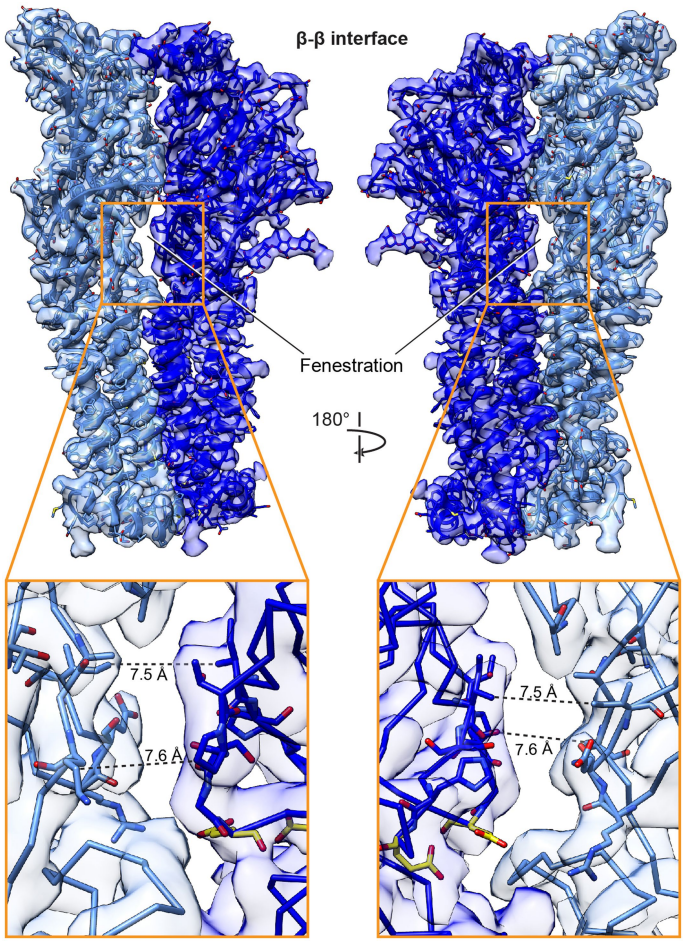


Extended Data Fig. 6 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

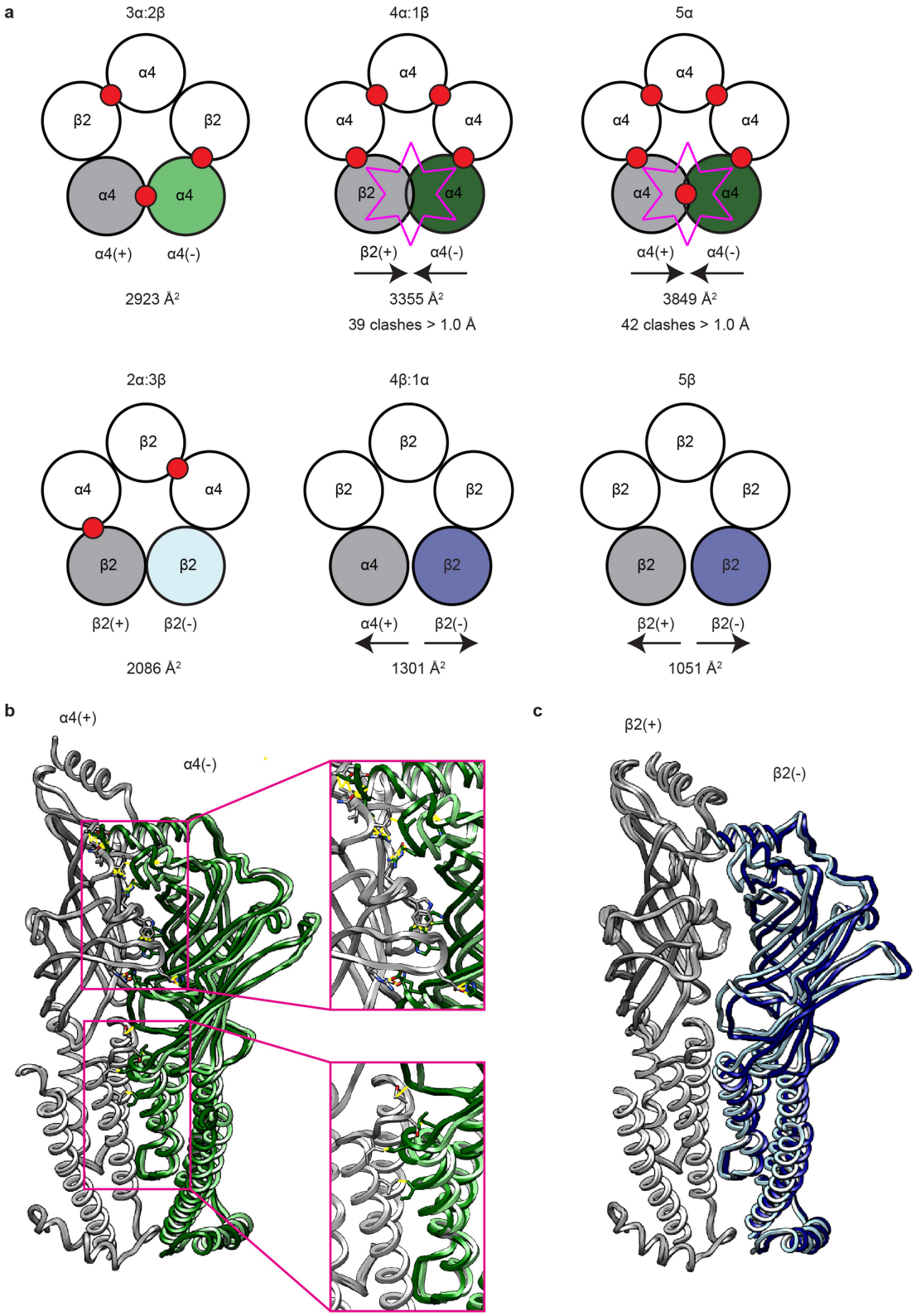
Extended Data Fig. 6 | Comparison of map and model of 2α3β by X-ray crystallography at 3.9 Å resolution and by cryo-EM at 3.4 Å resolution, and comparison of subunit backbone conformations. **a**, The X-ray electron density map (left) and cryo-EM map (right) are displayed in a linear representation (unwrapped) to illustrate the overall resolvability of all five subunits: two α4 subunits (green) and three β2 subunits (blue). Red circles indicate the locations of crystal contacts on the X-ray electron density map (left) or Fab binding sites on cryo-EM density map (right). **b**, X-ray structure of the 2α:3β assembly (PDB accession 5KXI[8]) after alignment to the cryo-EM structure (this study) shown in ribbon representation and coloured according to Cα r.m.s.d. values (from blue (low) to red (high)). While the areas involved in crystal contacts displayed relatively higher r.m.s.d. values for Cα (>2 Å), the Fab binding location shows r.m.s.d. values for Cα of <1 Å. This result implies that zones of protein–protein contacts in the crystal lattice may substantially affect local structure on the pentamer, but Fab binding in the cryo-EM study did not result in substantial structural change, consistent with the binding assays and electrophysiological results (Extended Data Fig. 1). **c**, Fit of X-ray model (5KXI) to X-ray electron density map (left) and cryo-EM model to cryo-EM density map (middle) shown as selected areas for high r.m.s.d. (loop C) and low r.m.s.d. (M2 helix), respectively. Right column shows superposition of X-ray and cryo-EM model at the selected areas. **d**–**h**, Subunit superpositions. **d**, Comparison of all α4 subunits in the 2α:3β and 3α:2β assemblies reveals no substantial conformational differences in α-carbon backbones. **e**, Superposition of all β2 subunits in the two assemblies reveals a domain motion of the TMD relative to the ECD of the subunits that comprise the β–β interface. The rotation of 4.4° between the two β-subunits that compose the β–β interface is denoted by black double-headed arrows. The principal β-subunit (β2 +) is coloured magenta while the complementary β-subunit (β2 −) is shown in cyan. **f**–**h**, Superposition of α4 (green) with the three distinct β2 conformations: principal β–β (magenta) complementary β–β (cyan) and α–β–α (grey). The conformation of the principal β–β (magenta) subunit is distinct from every other subunit in the 2α:3β and 3α:2β assemblies. The conformation of the complementary (−) subunit of the β–β interface is rotated towards the pore axis, accommodating the marked conformational change of the principal β2 subunit away from the pore axis. This conformational change results in the complementary β–β (cyan) subunit adopting a backbone conformation similar to α4 subunits.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



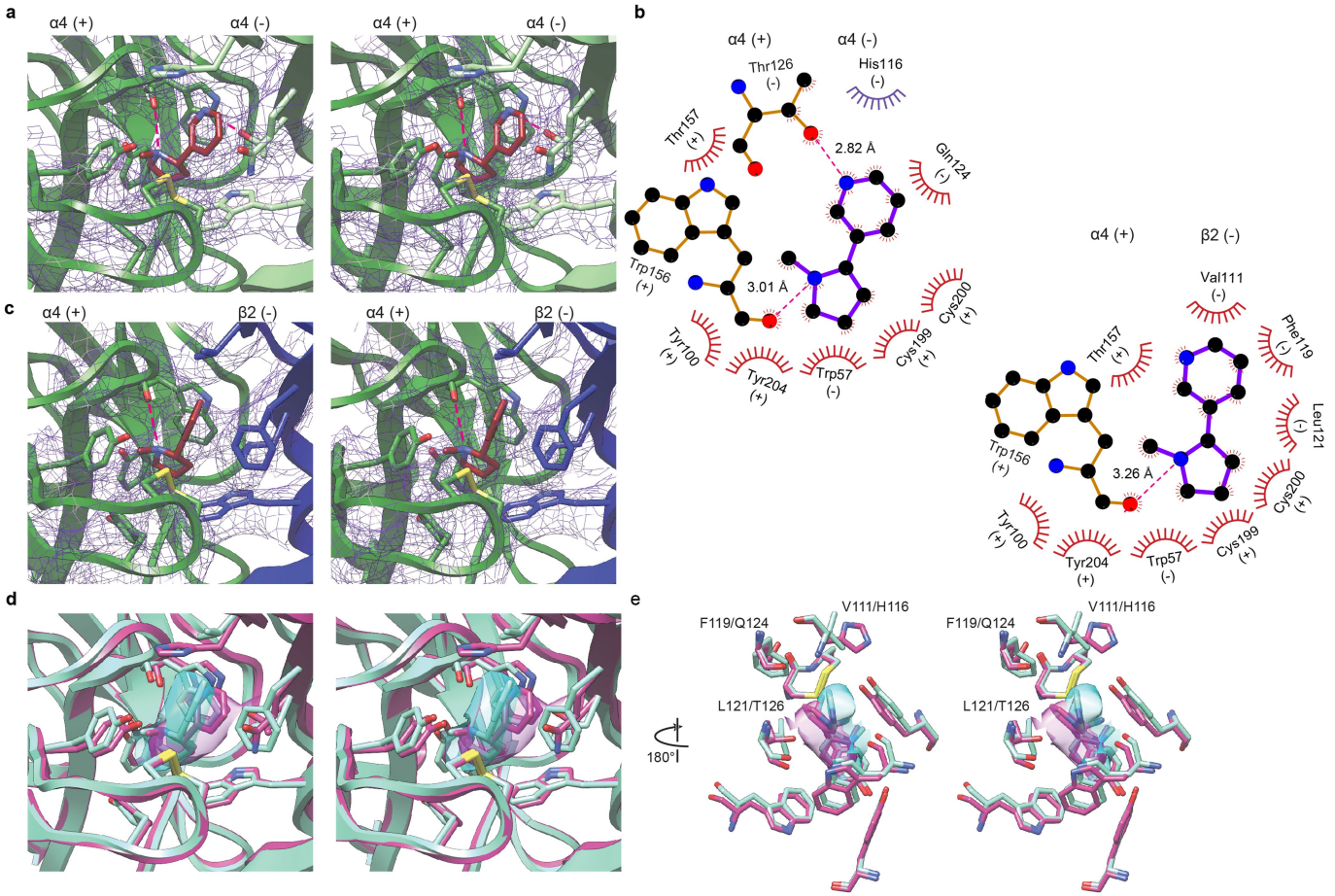
Extended Data Fig. 7 | Fenestration at β–β interface. Extracellular fenestration unique to this class of interface. Inset of fenestration with distances indicated by dashed lines and side chains surrounding fenestration shown as sticks. Two conserved glutamic acid residues in close proximity to this gap at the β–β interface (yellow) have been implicated in Ca^{2+} potentiation of nicotinic receptors[49].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



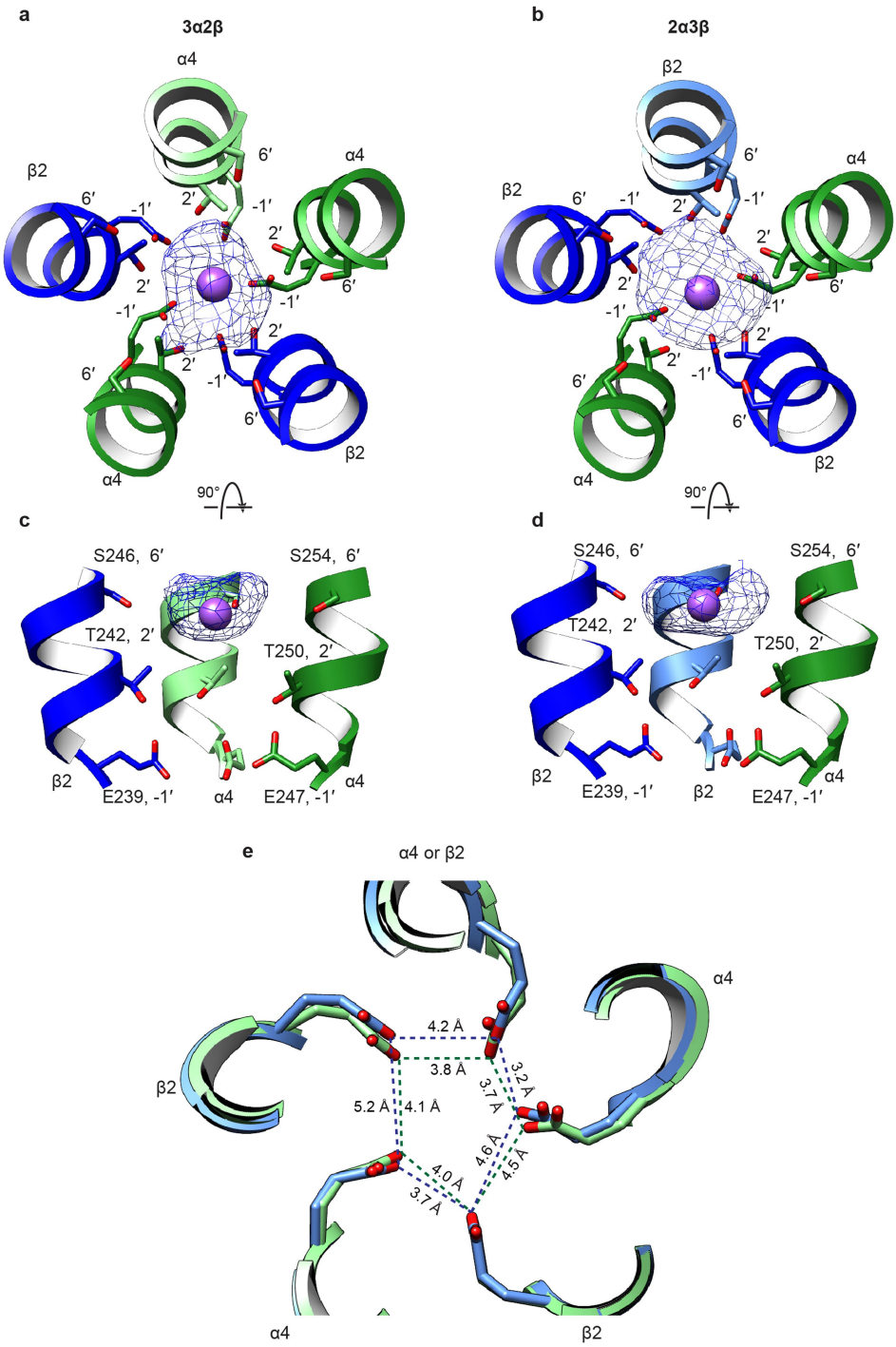
Extended Data Fig. 8 | Basis of α4β2 heteromeric assembly. **a**, Cartoon representation of top view of observed (2α:3β and 3α:2β) and computational (4α:1β, 1α:4β, 5α and 5β) α4β2 pentameric assemblies. Assemblies on the top row (3α:2β, 4α:1β, and 5α) are arranged by increasing α4 composition. Assemblies on the bottom row (2α:3β, 1α:4β and 5β) are arranged by increasing β2 composition. Agonist binding sites are denoted by red circles. Buried interface areas (Å^2) for the interfaces analysed in **b**, **c** are listed below each pentameric assembly. Subunits are coloured in **a** as described below for **b**, **c**. **b**, Superposition of α–α from 3α:2β and final α–α interface in 5α homopentamer. Principal subunits (grey) were superimposed to highlight differences at the interface. Complementary (−) subunits are coloured light green for α–α from 3α:2β and dark green for the α–α interface in the 5α homopentamer. Sticks are displayed for amino acid clashes that have greater than 1.5 Å overlap assessed by Molprobity. **c**, Superposition of the β–β interface from 2α:3β and final β–β interface in the 5β homopentamer. Principal subunits (grey) were superimposed to highlight differences at the interface. Complementary (−) subunits are coloured light blue for the β–β interface from the 2α:3β assembly and dark blue for the final β–β interface in the 5β homopentamer.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



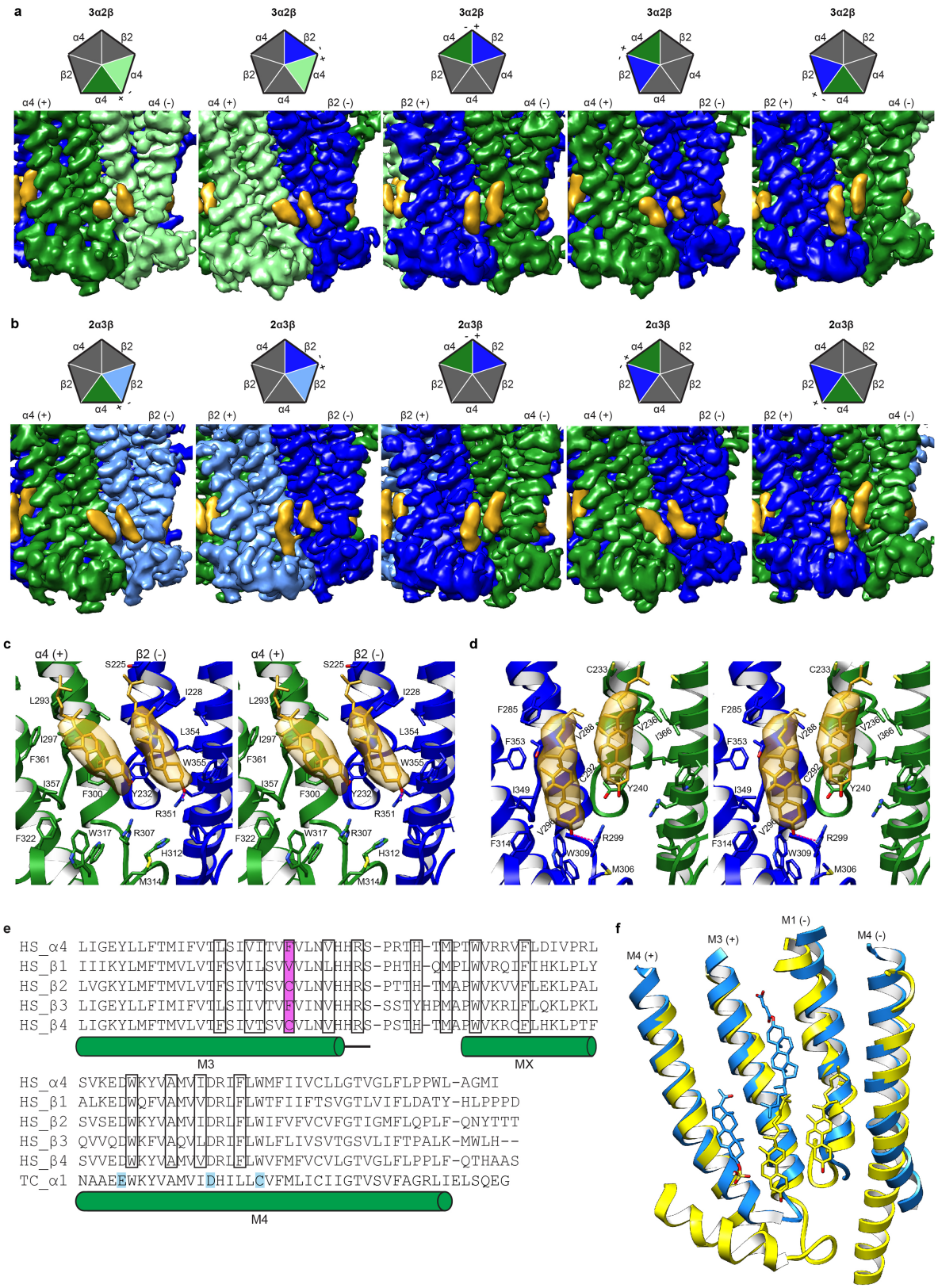
Extended Data Fig. 9 | Ligand binding site comparisons. a, Stereo images of nicotine bound at the α–α interface. Subunits are coloured as in Fig. 1. Density (purple mesh) is displayed at a threshold of 0.0306 in Chimera. Electrostatic interactions denoted as dashed magenta lines. Binding pocket residues and nicotine displayed as sticks. **b**, Ligplot of nicotine bound at the α–α and α–β interfaces (left and right, respectively). Residue contacts within 4.5 Å are displayed in red, residue displayed in purple (His116) is within 5.5 Å. **c**, Stereo images of nicotine bound at the α–β interface. Density (purple mesh) is displayed at a threshold of 0.0306 in Chimera. Hydrogen bonds denoted as dashed magenta lines. Binding pocket residues and nicotine displayed as sticks. **d**, **e**, Stereo image overlays of nicotine bound at the α–α (magenta) and α–β interfaces (cyan). Ligand density is displayed as transparent surfaces. Density is displayed at a threshold of 0.406 for the α–α (magenta) and 0.500 for the α–β (cyan) interface in Chimera. **e**, Ribbon removed for clarity, rotated 180° relative to **d**. Key residues implicated in the difference in sensitivity between the classical α–β neurotransmitter binding site (high sensitivity V111/F119/L121) and the unique α–α binding site (low sensitivity H116/Q124/T126) are labelled.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | Comparison of ions located in the pore above the constriction point. a, Top view of EM density map of ion in the pore of the 3α:2β assembly. Density (blue mesh) is displayed at a threshold of 0.027 in Chimera. Subunits are coloured as in Fig. 1. Modelled Na^+ ion is represented as a purple sphere. Nearest residues on the M2 α-helices are indicated. **b**, As **a**, but for the 2α:3β assembly. **c**, Side view of EM density map of ion in the pore of the 3α:2β assembly. Two subunits (one α and one β) removed for clarity. Colours and residues indicated as in **a**. **d**, As **c**, but for the 2α:3β assembly. **e**, Superposition of constriction region of 2α:3β (light blue) and 3α:2β (light green) assemblies.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 11 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 11 | Comparison of putative cholesterol binding orientations. a, **b**, EM density map showing cholesterol sites from all TMD interfaces in the 3α:2β assembly and 2α:3β assembly, respectively. Pentagons (top) are coloured to illustrate subunits composing the displayed interface. **c**, Stereo image of representative cholesterol binding site with α-subunit on principal side (+). When the principal side (+) is an α4 subunit, both molecules tilt towards the principal face. One residue from the complementary side β2 (−), Y232, contacts both cholesterol molecules at the α–β interface (at the α–α interface the equivalent residue Y240 in the complementary (−) α4 contacts both cholesterol molecules at the interface; not shown). **d**, Stereo image of representative cholesterol binding site with β-subunit on principal side (+). When the principal side (+) is a β2 subunit, both cholesterol molecules are oriented orthogonal to the plane of the membrane. One residue from the principal (+) β2 subunit, V288, makes contacts with both cholesterol molecules at the β–α and β–β (not shown) interfaces. Cholesterol and interacting side chains are shown as sticks. Density maps in panels **a**–**d** are displayed at a threshold of 0.025 in Chimera. **e**, Alignment of nicotinic receptor subunits for region encompassing putative cholesterol sites. Residues in cholesterol binding pocket are boxed and amino acid position implicated in differential cholesterol binding is highlighted in magenta. Uniprot accession IDs are provided[40]. Residues mapped using a photoreactive cholesterol analogue in *Torpedo californica* (TC P02710)[26,27] are highlighted in cyan. All other sequences are from *Homo sapiens (*HS P43681, P11230, P17787, Q05901, P30926 for α4 and β1–4, respectively*)*. **f**, Superposition of β–α (yellow) from $\alpha_4\beta_2$ and α1–α1 (blue) $GABA_A$–GLIC chimaera (PDB accession 5OSC[25]) TMD interface to compare putative binding sites. Pregnenolone sulfate, CHS and cholesterol are shown as sticks.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

Dataset	2α3β stoichiometry EMDB-7535 PDB-6CNJ	3α2β stoichiometry EMDB-7536 PDB-6CNK
Data collection and processing		
Magnification	46,730	
Voltage (kV)	300	
Electron exposure (e-/Å^2)	50 (2,588 images) or 75 (3,323 images)	
Defocus range (μm)	-0.5 to -4.5	
Pixel size (Å)	1.07	
Symmetry Imposed	C1	
Initial number of particles	649,773	
Final number of particles	263,695	139,551
Map resolution (Å)	3.4/3.7*	3.7/3.9*
FSC threshold	0.143	0.143
Refinement		
Intial model used (PDB code)	5KXI	5KXI
Model resolution (Å)	4.2	4.3
FSC threshold	0.5	0.5
Map-sharpening *B* factor (Å^2)	-150	-170
Model composition		
Non-H atoms	25,347	22,082
Protein residues	3,183	2,757
B factors (Å^2)		
Receptor	128	115
Fab	233	201
Glycosylation	135	117
Ligand	107	101
Cholesterol	142	140
Ions	117	27
Validation		
MolProbity score	1.9 (100th percentile)	1.98 (100th percentile)
Clashscore	11.8	13.6
Poor rotamers (%)	0.25	0.16
R.m.s. deviations		
Bond lengths (Å)	0.005	0.007
Bond angles (°)	0.68	0.69
Ramachandran plot		
Favored (%)	95.5	95.1
Allowed (%)	4.5	4.9
Disallowed (%)	0	0

*Mask including Fab

*Mask including Fab.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

CAREERS

PERPSPECTIVE How creative writing can inspire your research **p.269**

BLOG Personal stories and careers counsel **http://blogs.nature.com/naturejobs**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

ILLUSTRATION BY NEIL WEBB/DEBUT ART



MENTAL HEALTH

Out of the darkness

Depression and related disorders can be devastating, but there are ways to fight them.

Mental illness is widespread in the sciences, and graduate students are particularly vulnerable (*Nature* **556**, 5; 2018). Building a strong support network and separating your sense of self from your job are crucial. Here, in the second of our three-part series on mental health in academia, four researchers share their stories and advice. Next week, we examine lab leadership and and how to improve the health of research groups.

NATURE.COM
Read more on mental health in science at go.nature.com/wellbeing

ELYN SAKS
'Crazy thoughts'

Lawyer and psychoanalyst at the University of Southern California Gould School of Law, Los Angeles

It was 1982, and I was in my first semester as a student at Yale Law School. We had a paper due and I climbed onto a roof, singing, dancing, gesticulating and saying crazy things, like I had killed people with my thoughts. My friends called the student health centre and tried to get me there, but I wouldn't go.

The next morning, I asked for an extension on the paper. I was still saying crazy things, and my professor took me to the emergency department. They committed me. I was hospitalized for five months, restrained, forcibly medicated, given very little privacy and locked up.

If you were withdrawn from Yale for psychiatric reasons, you had to be evaluated for readmission by the head of university health ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

psychiatry. I looked this person up and found an article he'd written on exactly this topic — the questions that the evaluator should ask, and the answers to look for. I was totally prepared, and it unfolded exactly as he had laid it out.

Still, part of me felt like I would never get back on my feet. The guy advised me to become a cashier for two or three years. I asked myself how much more stressful it would be to stand at a cash register dealing with a long line of people waiting for change from their purchase. So I went back to Yale.

I was eventually diagnosed with schizophrenia, and it took me ten years to become reconciled to staying on medication. Once I did, my life got much better. In a work environment, it also helps if you have people who know what you're going through. But some people will think you're not up to the job or are dangerous or scary. You've got to pick your supporters carefully.

And stigma is a real scourge. When my memoir, *The Center Cannot Hold: My Journey Through Madness*, came out in 2007, an administrator on the staff at the University of Southern California Gould School of Law told me that she was glad she didn't know I had schizophrenia when we started going to dinner together. She said she'd never have gone. I was stunned that a smart, kind, well-meaning person would have such a picture of mental illness that she wouldn't have gone to dinner with me.

I've learnt that my mind is my best friend and my worst enemy. When I'm working, the crazy thoughts recede.

NATHANIEL BORENSTEIN
'My life sucked'

Veteran computer scientist and Internet pioneer in Greenbush, Michigan, who helped to create e-mail and the precursor to PayPal.

When I was in graduate school in the 1980s, studying the theory of computation, my adviser — who had been very encouraging — suddenly started indicating that he didn't understand what I was doing. I got very depressed and almost left the programme. I didn't perceive my reaction as symptomatic of an illness. I just thought that the whole world sucked, that my life sucked and that everything was wrong.

I get depressed a couple of times a decade, and my depressions tend to look the same. I spend a lot of time curled up in bed, not wanting to talk to anyone. Everything looks bad.

I think that the better you know your pattern, the more likely you are, and the quicker you are, to recognize it. When I start to recognize my pattern, it's comforting to know that the world is not completely horrible and that I'm just depressed. I immediately make an appointment with my therapist. I want to be open about this, so that when younger people read this, they can be spared that response. At least if they know it's an illness, when they're curled up in a ball, it's better than thinking that the whole world sucks.

There is a tremendous resistance to taking psychiatric drugs, but I would encourage people who are struggling with depression to consider them. In the early 1990s, I started antidepressants and that was a godsend. Within a couple of hours, I was feeling better. And a few days after that, I met someone I totally clicked with professionally and we founded a new digital company, the predecessor to PayPal.

The times that you might think were my biggest triumphs, when you might have expected me to be dancing around and doing a jig — those are the times I've tended to get extremely depressed. It's because I have accomplished this thing and it was a lot of work — but it was nowhere near what I thought I ought to have accomplished. I choose one path and look at someone who has gone down a different path, and I feel like I've fallen short.

"I choose one path and look at someone who has gone down a different path, and I feel like I've fallen short."

My most recent episode was in January and February of this year. I went to the hospital with constant, extreme abdominal pain, which was making me very depressed, and was diagnosed with colitis. That was the first time I'd thought of suicide as a way to end the pain. Since then, I have got help and the pain is under control.

I've heard that depression is a gift, because it can help us to recognize when we need to change something. But if anyone had told me in 1981 that my depression was a gift, I would not have believed it for an instant. I once came this close to destroying my career.

ANONYMOUS
'Know your limits'

Male molecular biologist in the United States.

I was in graduate school in South America. I was frustrated there and had a period of depression. My principal investigator (PI) suggested I see a psychiatrist, who gave me medication for bipolar disorder, and it helped. I also realized that graduate school would be over soon and that I'd be able to move on to the next thing.

In January 2015, about six months before graduating, I started looking for a postdoc position and researched a lot of labs. I sent my CV to 40 or 50 places, along with detailed cover letters. And I had a really good publication record. But I was not getting any responses. I became despondent, fell into a debilitating depression and began drinking heavily. It was a black hole.

The whole time I was applying, my PI kept saying: "Why don't you write a postdoc fellowship to stay in the lab here with me, in case things fall through?" So I did. And my fellowship application came through with high marks and got funded. But I didn't stop looking for other postdocs, and began applying to a couple of places I didn't tell my PI about.

One didn't ask for a recommendation letter, and when the people there saw my CV, they hired me on the spot. Only later did I discover that my PI had been sabotaging my career.

Academia has structural disadvantages for people with mental-health issues. To a great extent, your career path is decided by factors beyond your control. Confidential letters of recommendation are very important in the hiring process, particularly in science. You're not going to know that someone threw your CV in the garbage because someone else said, "Hey, he's depressed". Those opportunities are lost for ever.

Unless there is a structural change in the system, that bias is not going away. My advice? Know your rights. Carefully document any instances of abuse or misconduct. And try to ensure that a former employer will have nothing to use against you that could affect your future career. This includes keeping your mental-health issues out of the workplace as much as possible, even if you have to keep odd hours in the lab to work around therapy.

Learn to carefully balance the needs of your job against your personal and mental needs. I take a mood stabilizer and can't drink alcohol at all. And I know when I need to leave work. I think, "I've spent enough time here. I think I'll swim in the pool." You have to know your limits.

ANONYMOUS
'You're not alone'

Female bioscientist in Europe.

When I applied to PhD programmes, I was accepted by two in the United States. One offered me funding and the other didn't. But the one that didn't was more prestigious, and I was misled into thinking that I would have no trouble getting teaching-assistant positions to help pay for the programme. So I decided to go there. I made a very big mistake.

I arrived in the summer and got a position as a teaching assistant — but I had to reapply for a new one every couple of months. On top of that, I got accepted as a research assistant in the lab of a famous professor, but he paid nothing. I was also taking a full load of courses and studying for qualifying exams.

The first semester was very difficult. I couldn't focus. I was always lost. I was always sad. I had to drop a course. If only someone

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

at that time had said: "You have depression and that's why you can't focus."

The next semester, I took a lighter load. I thought it would be easy. But I got a low score. I also wrote an internal grant that got funded, but I never saw the money. At that point, I didn't want to get up or out of bed. I started being absent from the lab. I couldn't bring myself to walk to that place. The worst part is that I ended up with a C. My grade-point average crashed. Then, when exams came, I failed miserably. I worried about my teaching-assistant position and losing that money. I panicked. I was sitting there crying and thinking my world had fallen apart.

GETTY

I called a counsellor and told them that I was considering self-harm. There is a law in that area that if you try to hurt yourself, they put you in the psychiatric hospital. The police came in and handcuffed me. I was just crying and repeating, over and over, "I got a C. I'm going to get kicked out." In the hospital, I was monitored 24/7. After two days, I was still crying.

Ten days after my first hospitalization, I was handcuffed again and put in the hospital again. Each time I was put in the hospital by the police, I was charged more than US$100 — a huge expense for me. I contacted the other PhD programme to see if the people there would honour their funding offer from the previous year. They said no. I opened a bottle of pain medication and started popping the pills in my mouth. Then I stopped and called my psychologist. After a few weeks off from my PhD studies, another hospitalization and two failed exams, my position in the programme was terminated.

"Reach out for help. Surround yourself with people who will show you, in this darkness, that you will make it."

I started looking for other PhD courses and found one with a renowned scientist in Europe that would pay me enough, so I went there. But you can't just say, "It worked out in the end." For a year, I sat and stared at the wall. Even though the position is well paid, the culture still treats PhDs poorly and values unhealthy competition. I am leaving academia.

Tell the darkness not to take you today — not this moment, not this second. Take time off if you need to. Find the person you were before the darkness came. Reach out for help — surround yourself with people who will show you, in this darkness, that you will make it. As isolating and hopeless as it might seem right now, you are not the only one who has had to pass through this. You are not alone.

INTERVIEWS BY EMILY SOHN

Interviews have been edited for clarity and length.

COLUMN

Tell the stories in your science

Use principles of narrative fiction to inspire your research papers, says **Amanda C. Niehaus**.



In the final month of my Australian Research Council fellowship at the University of Queensland in Brisbane, Australia, I published papers about sex-crazed marsupials, wrote grant applications and finished *The Breeding Season*, an as-yet-unpublished novel.

Yes, I wrote a novel — fiction founded on my science research. The story takes what I know about reproductive conflict in wild animals and explores it in humans. It's the hardest and most gratifying thing I have ever done.

I never expected to write about science in this way, mainly because I thought science and art were mutually exclusive. I learnt to avoid bias and passion, and to present my data as if seen through a microscope from a careful distance. Science is objective. Art, by contrast, is all about perspective, subjectivity and the confused experience that is life.

But creative writing can make scientific communication more powerful. We remember stories because our brains are wired to: we find them more interesting and are more likely to retell information if it's presented as narrative rather than exposition. In fact, papers written in a more-narrative style might be published in higher-impact journals and cited more often (A. Hillier *et al. PLoS ONE* **11**, e0167983; 2016).

I've spent the past five years collecting data on ageing in northern quolls (*Dasyurus hallucatus*), small marsupials whose males essentially copulate themselves to death. As a scientist, I analyse, interpret and present the data with objectivity, according to the conventions of my field.

But I don't want to put distance between myself and my work, or between my work and the public. I want to show that the questions these animals face are universal — for example, how much time and energy to expend in making babies, and what that investment means for every other aspect of life. These issues define the choices we make about when and whether to have children, how to balance our careers with our families and how to spend our finite lives.

This is messy stuff. It's complex and complicated in ways not easily explored in a paper.

Last year, I published an essay called 'Pluripotent' (A. C. Niehaus *Creative Nonfiction* **64**, 22–26; 2017), which examines the role of stem cells in the context of my experiences as a mother, researcher and cancer survivor, but also, metaphorically, as a woman in science. Creative writing can bring difficult concepts to life.

For me, compressing science into academic journals simply isn't enough. I'm frustrated by the need to reduce my ideas and experiments to publishable pieces while simultaneously ensuring that they are broadly relevant. I believe that the most exciting things happen at the fringe, the overlap, the moment we look at the same question through a different lens altogether. New ideas happen outside our comfort zone.

My research shows that among northern quolls, males and females lead different lives because of variations in how their bodies balance reproduction and longevity. How would these differences affect a reproductive relationship between humans? Their careers? Their ambitions? Isn't this what many of us really want to know?

In writing quoll biology into my novel and a short story, I discovered that artists and writers seek truth as much as scientists do. They embed facts with experiences to give them context and meaning. And stories deal not only with what is true, but also with what is possible. Through fiction, I may discover something about sex and death that my research did not tell me.

Where to start? Take a workshop in creative writing, curate an online gallery of inspiring images or invite writers to your next symposium. Stories are there in every book, movie and conversation — so notice them, and harness their energy to share your work. ■

Amanda C. Niehaus *is a scientist and writer in Brisbane, Australia.*

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

BLUE SKY RESEARCH

Discovery can't afford to knap.

BY IAN STEWART



ILLUSTRATION BY JACEY

The sun beat down like a hammer.

Stoneseeker scoured the base of the cliff, looking for rock samples to take back to the elders. It was his first solo exploration, and his brief was simple: *find more flint*. The chief had relieved him of hunting duties, providing food and a waterskin. In return, the young man had promised to seek out a new source of the dense black flint that could be knapped into knives, arrowheads and spearheads, for the old sources were running out.

He tracked speckled seams of dark stones as they wove along white cliffs, but the quality of the flint started out poor and steadily got worse. Eventually the seams ran out altogether, mainly because the cliffs ran out. He could not return to the tribe empty-handed, not when it had provided so generously for his expedition. He camped for the night, making a fire to scare away predators.

Next morning, he walked across shattered rock and gravel, heading for a distant hill. If not flint, perhaps there would be obsidian! The glassy material was hard to work, but its shards were sharp and the edge lasted longer.

No flint, no obsidian. But he did find an outcrop of rock, of a kind he had never seen before; almost the colour of the sky, but tinged with green. *Interesting*. Curiosity aroused, he put a few lumps in his shoulder-bag.

Flint, he reminded himself, and walked on. Two days later he gave up, and turned back to face the tribe's wrath.

The chief towered over him, splendid in eagle-feather robes. Stoneseeker trembled. "I failed, Bearkiller," he confessed. "I found no flint."

"That was not your promise when you were relieved of mammoth-hunting duties and given food caught by others!"

I promised to look, not to find, Stoneseeker thought. But it was unwise to contradict the chief.

"Don't hunt, don't eat," Snakebite the shaman jeered.

"I did find *these*," Stoneseeker said humbly. Bearkiller took the strange bluish stones, and passed them to Blademaker. The flint-knapper placed one on his thigh-pad and struck it with a hammer-stone. It crumbled into fragments.

"Won't take an edge," he complained. "Useless!"

Stoneseeker, still intrigued by the strange stones, gave utterance to a thought that had been growing in his mind. "Bearkiller, I seek a boon."

"A boon? For failure?"

"I am Stoneseeker! I seek stones. I *found* stones. Er — just not flint."

"You found *wrong* stones!" roared Bearkiller. "*I sent you for flint!*"

"Well, yes … but I found *these* stones instead, great chief, and I'm wondering whether it's worth investigating them. They have the colour of the sky. Such marvels could have powers undreamt of!"

Blademaker sneered. "Can your sky-stones make a spear? A knife? Even a rudimentary hide-scraper?" He picked up the fragments, let them fall through his fingers. "They don't even make serviceable hammer-stones!" He spat. "Useless!"

"How do you plan to investigate the stones?" Bearkiller asked quietly.

"Uh — I haven't really got that far, great chief. Drop them in water? Throw them at hares? Leave one lying in the moonlight? Rub them with —"

"And what do you propose to find?"

"I have no idea."

"Then what *use* are they?"

Stoneseeker sought a helpful analogy. "Great chief: what use is a newborn baby?"

"Not my department. Call the midwife."

"Don't listen to the idiot!" Mammoth-stabber yelled. "He's just trying to con us into feeding him while he weasels out of the hunt!"

"Don't hunt, don't eat," the shaman repeated.

Bearkiller stroked his impressive beard thoughtfully. "Half a day's walk from here is a tar-pit," he said enigmatically. "What do we find there, Stoneseeker?"

"Lions," Stoneseeker mumbled. He could see where this was going. "Dead ones."

"And what made them walk into the tar to die?"

The real answer was trapped prey, for food, but what the chief expected was the conventional folk-wisdom.

"Curiosity," Stoneseeker whispered.

"*Curiosity killed the cat!*" the shaman declaimed.

"You mean well, youngster," said the chief. "But the tribe can't squander precious food to let you waste good hunting-time, merely to satisfy your curiosity." He raised his eyes to the heavens. "We need *practical* ideas. As well as try to find out why your stones are the colour of the sky!" The other elders howled with laughter. "You can't tell us what you will investigate. You can't tell us what you will find. You don't know what you're doing."

If I knew beforehand what I'd discover, there'd be no point in trying to discover it, Stoneseeker thought, but remained silent.

Stoneseeker showed the sky-stones to Foxfoot, who made pots. "Heat does strange things to colours," she mused. "If I grind some of these they might just work as glaze. Or the heat might make the stones stronger. I've built a new kiln, the hottest yet! I'll heat them overnight."

Next morning, when Foxfoot looked for the stones in the kiln, there was nothing but grey ash. Raking it out, something strange caught her eye. She picked it up and wiped away the ash. It was a lumpy tangle of reddish material. Some bits glistened in the sunlight.

They took it to show Bearkiller.

"I told you not to investigate further."

"We did it in our spare time, great chief."

"People who eat at the tribe's expense but produce nothing should not have spare time." He glared at Stoneseeker. "You're going to ask me to reverse my decision, aren't you? Just because you've found something you don't understand." He turned the irregular mass of — *something* — in his hand. "I'll ask the expert. *Blademaker!*"

The flint-knapper rushed over.

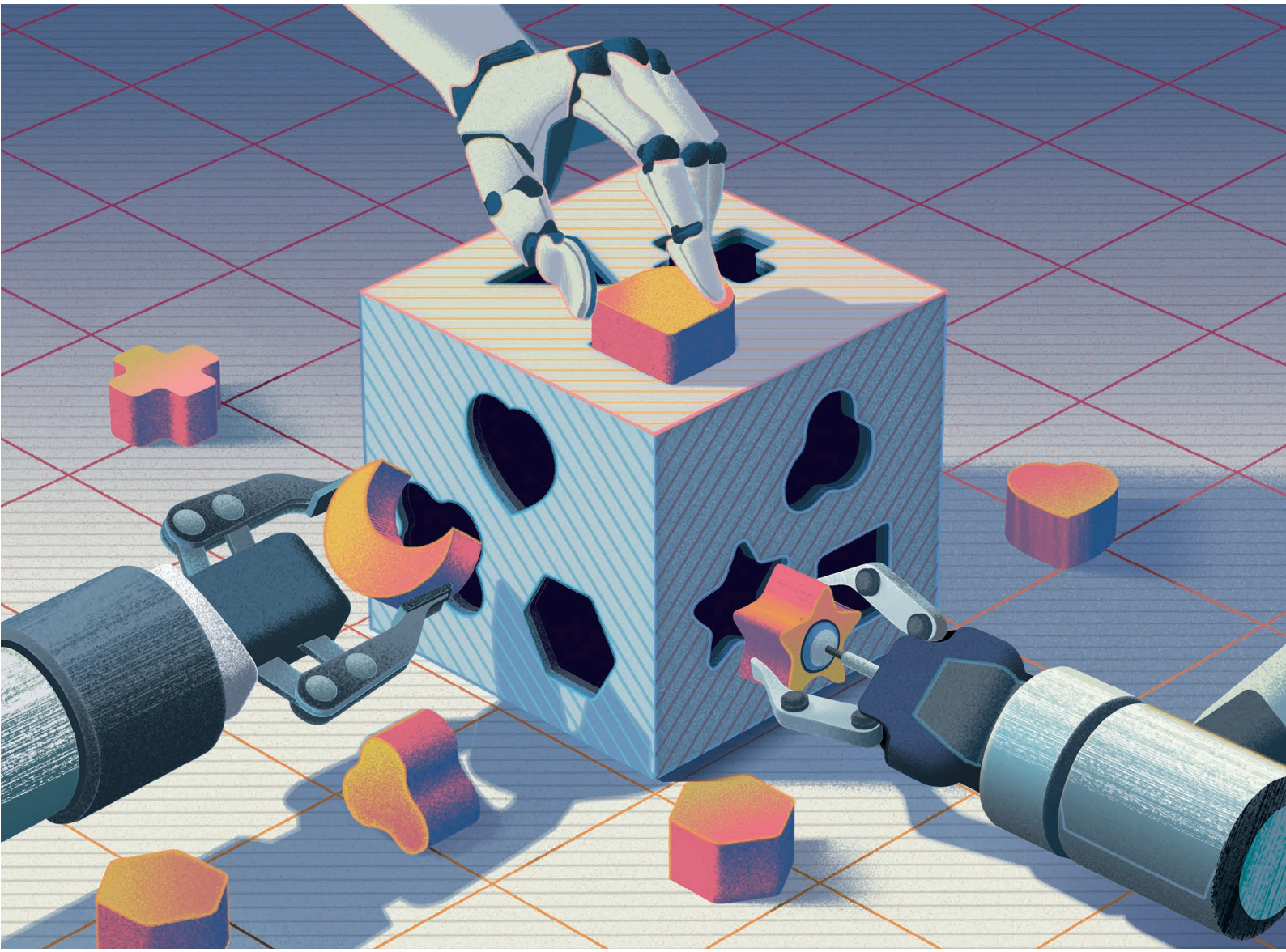
"What do you think of this stuff?"

Blademaker took the strange lumpy material and grunted unhappily. It bent easily in his hands. "Too soft. It will never replace the flint spearhead! Useless!"

Bearkiller nodded. And threw it away. ■

Ian Stewart's *latest SF novel is* Rock Star. *He is sole winner of the Bloody Stupid Johnson award for Innovative Uses of Mathematics, and an honorary wizard of Unseen University.*

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



A gripping problem

Designing machines that can grasp and manipulate objects with anything approaching human levels of dexterity is first on the to-do list for robotics.

BY RICHARD HODSON

ILLUSTRATION BY SAM CHIVERS

In the 1973 science-fiction classic *Westworld*, the robots that staff the film's fictitious holiday resort are indistinguishable from their human guests, except for one small clue: the engineers, we are told, haven't perfected the hands yet.

The capabilities of real-world robots fall a long way short of those of *Westworld*'s murderous hosts, but on this small point, reality and fiction are in agreement: hands, and the manipulation of objects, are particularly challenging aspects of robotics. "Grasping is the critical grand challenge right now," says Ken Goldberg, an engineer at the University of California, Berkeley.

In the past 50 years, robots have become very good at working in tightly controlled conditions, such as on car-assembly lines. "You can build a robotic system for one specific task — picking up a car part, for example," says Juxi Leitner, a robotics researcher at the Australian Centre for Robotic Vision (ACRV), based at the Queensland University of Technology in Brisbane. "You know exactly where the part is going to be and where the arm needs to be," he says, because the robot picked up the same thing from the same place "the last million times".

But the world is not a predictable assembly line. Although humans might find interacting with the countless objects and environments found beyond the factory gates a trivial task, it is tremendously difficult for robots. ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

These unstructured environments represent the next frontier for researchers across the field of robotics, but such environments are particularly problematic for robots that grip. Any robot hoping to interact physically with the outside world faces an inherent uncertainty in how objects will react to touch. "We can predict the motion of an asteroid a million miles away far better than we can predict the motion of a simple object being pushed across the table," Goldberg says.

Some researchers are using machine learning to empower robots to independently identify and work out how to grab objects. Others are improving the hardware, with grippers ranging from pincer-like appendages to human-like hands. And roboticists are also gearing up to tackle the challenge of manipulating objects gripped in the hand.

Advances in robots' ability to handle objects could have enormous societal impact. Commercial entities, particularly those involved in the movement of varied goods, are following developments closely. "There's a big demand. Industry really wants to address this because of how fast e-commerce is growing," says Goldberg. With interest greater now than ever, "It's an opportunity for the research to really be put into practice."

LEARNING TO LEARN

The heightened industry interest is exemplified by an annual competition organized by e-commerce giant Amazon for the past three years. The Amazon Robotics Challenge asks teams of researchers to design and build a robot that can sort the items for a customer's order from containers and place them together in boxes. The items are varied, ranging from bottles and bowls to soft toys and sponges, and are initially jumbled together, which makes it a difficult task in terms of both object identification and mechanical grasping.

In July 2017, Leitner's ACRV team claimed victory with a robot called Cartman, which resembles a fairground 'claw' game. An aluminium frame supports the claw assembly. The robot has two tools for picking up objects, known as end effectors: a gripper with two parallel plates, and a suction cup backed by a vacuum pump. For each object the robot encounters, the researchers specify which effector it should try first. If that doesn't work, the robot switches tools.

First, however, the robot must find the item it's looking for. The team tackled this challenge by using machine learning. The main input for the robot is an RGB-D camera, a technology that is popular among roboticists and that can assess both colour and depth. The camera looks down from the effector into the boxes below. From this vantage point, Leitner explains, Cartman labels each pixel according to the object it belongs to — a form of deep learning known as semantic segmentation. Once a cluster of pixels representing the desired object is found, the camera's depth-sensing capability helps the robot to work out how to grab the item. "In simple terms, we attach to the bit that pokes out the most," says Leitner.

Rapid advances in machine learning underpin a lot of recent progress in grasping. "Software has been the bottleneck for ages, but it's becoming more advanced thanks to deep learning," says Pieter Abbeel, a deep-learning specialist at the University of California, Berkeley. These developments have, he says, opened up "whole new avenues of robotics applications".

Abbeel is co-founder and chief scientist of covariant.ai, a start-up in Emeryville, California, that uses deep learning to train robots. Rather than program a robot to perform a specific action, humans provide demonstrations that the robot can then adapt to deal with variations of the same problem.

> "GRASPING IS THE CRITICAL GRAND CHALLENGE RIGHT NOW."

The human trainer views the camera feed of the robot arm through a headset, and uses motion controls to guide the robot arm to pick up objects. The process feeds a neural network with data on the approach taken. "With just a few hundred demonstrations done in this particular way, you can train a deep neural network to acquire a skill," says Abbeel. "And I don't mean acquire a specific motion that it's going to repeatedly execute, but acquire the ability to adapt the motion to whatever it's seeing in its camera feed."

Goldberg uses machine learning to teach robots to grasp, too. But rather than gather data from real-world attempts, his Dex-Net software is trained virtually. "We can simulate millions of grasps very quickly," he says. The software lets an industrial robot pick objects from a pile with a success rate of more than 90%, even if it hasn't seen those objects before. It can also decide for itself whether to use a parallel-jaw gripper or suction tool for a particular object.

Dex-Net's fourth incarnation will be presented in 2018. According to a metric being developed by Goldberg and roboticists around the world to aid reproducibility, known as mean picks per hour, the Dex-Net system is now among the fastest pickers around. It can achieve over 200 picks an hour — still behind human capacity, estimated at 400–600 picks an hour, but far ahead of the numbers achieved by the teams at the most recent Amazon Robotics Challenge (see 'A measure of success').

But Dex-Net's simulated world is imperfect. The model assumes that objects are rigid, for instance, and does not account for objects that contain liquid. Simulation, Abbeel says, might not always be the easiest way to learn. "The real world provides a simulator for free," he says. "It's great to leverage both ways of doing it."

SOFTLY DOES IT

The combination of a parallel-jaw gripper and suction employed by both Goldberg and Leitner is a popular choice: most teams at the 2017 Amazon Robotics Challenge took this hybrid approach. But how these tools grasp objects — planning contact points before moving into position as precisely as possible — is very different to how we humans use our hands.

"When you pick something like a pen up off a table, the first thing you touch is the table," says Oliver Brock, a roboticist at the Technical University of Berlin. We do not think about where we need to place our fingers. The softness of human hands allows for something called compliant contact — the fingers mould against the surface of the object. "Because you get a lot of surface contact, you can much more intuitively reach out and grab," says Daniela Rus, a roboticist at the Massachusetts Institute of Technology in Cambridge. "With soft fingers, we change the paradigm of grasping."

Many are seeking to exploit the benefits of compliance in grasping by building softness into robot grippers. Brock's lab has developed the RBO Hand 2, a human-like hand with five silicone fingers. The fingers are controlled by the movement of pressurized air, which allows them to curl and straighten as required.

Although the human-like arrangement of the fingers might not be suited to every task, it is ideal for interacting with the world we inhabit. "The world is designed for anthropomorphic hands," says Brock. But there's also a romantic element to the anthropomorphic design of his robotic hand. "It's embarrassing," he says. "People, even roboticists, are more fascinated by things that look human."

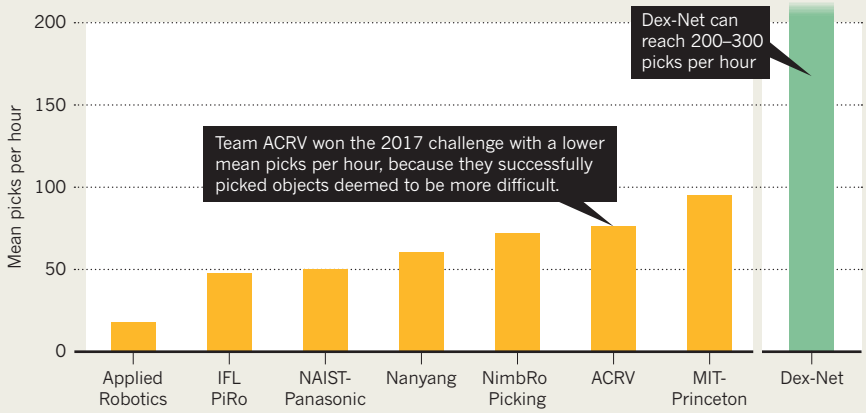
The benefits of softness are already drawing commercial interest. Soft Robotics, in Cambridge, Massachusetts, produces air-actuated grippers that have a more claw-like design than Brock's research model. The robots are already being trialled in a factory setting, handling delicate produce without damaging it.

Another start-up, Righthand Robotics in Somerville, Massachusetts, is adding softness to the claw and suction set-up so popular with roboticists. Its claws have three flexible fingers, arranged around a central suction cup that can be extended to draw in objects. The design takes inspiration from birds of prey, says the company's co-founder Lael Odhner. In these birds, most of the forearm musculature is attached to a single group of tendons that reaches to the tip of the claw, he says. Similarly, all the power of the motors in Odhner's robotic claws is put into a single closing motion. This

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

A MEASURE OF SUCCESS

To make it easier to compare experimental results in robotic grasping, Ken Goldberg, an engineer at the University of California, Berkeley, and others are preparing a metric known as mean picks per hour that could be used as a standard measure across the field. By this metric, the Dex-Net system seems to be considerably more capable than the robots that competed in the 2017 Amazon Robotics Challenge. They are all, however, still far behind humans, who can manage 400–600 picks per hour.



SOURCE: KEN GOLDBERG

simple action improves reliability — a crucial consideration for grippers intended for commercial use — at the expense of the ability to perform delicate motions. The extendable suction cup makes up for this shortfall. "It replaces potentially dozens of fine actuators that you would otherwise have to place in the hand," says Odhner.

IN OUR HANDS

Softness is still relatively new to robotics. "It's an extraordinarily powerful concept, but people are just beginning to figure it out," says Rus.

A common criticism is that it's hard for a soft robotic hand to perform a useful action with a grasped object. "You grab it very well," explains Rus, but "you don't know exactly the orientation of the object inside the hand". This makes it tricky to manipulate the object. Goldberg echoes this: "If you just have a soft, enveloping type of hand, then you've really reduced your visibility of that object," he says.

The way people solve this is simple: we touch. But few robots have been granted this capability. "People agree that it is important," says Brock — it's just very difficult to do. He is pursuing two approaches to giving his soft robot hands a sense of touch. The more-mature method involves embedding tubes of liquid metal in a silicone sheet wrapped around the finger. His team can then monitor applied forces using the electrical resistance along the tubes. "It's measuring strain all over the finger and inferring from that, through machine learning, what actually happens to the finger." The team is currently assessing how many of these strain sensors are required on each finger to measure various forces.

Brock's other approach to incorporating touch centres on acoustics. In a proof-of-concept test, a microphone was placed inside the air chamber of a soft finger. The sound that was recorded enabled researchers to identify which part of the finger was touching something, the force of the touch and the material of the object. The ability to tuck the microphone deep inside the finger — and therefore avoid reducing compliance — sidesteps a key issue with sensorizing soft fingers. Brock says that the details of this work will be published soon, and that he plans to work with acoustics specialists to improve his design.

Many roboticists think that there is unlikely to be a universal solution to grasping. Even if robots could achieve human levels of dexterity, Rus points out, "there are lots of things that we cannot pick up with a human hand." But as robots become increasingly adept at handling variability, more tasks currently performed by humans will become automatable.

A report published last year estimates that most occupations could be at least partially automated (go.nature.com/2rjtnud). But Goldberg stresses that robotics does not have to put people out of work. "My goal is not to replace humans," he says. "What I want to do is assist humans."

Whatever the outcome, a lot of development needs to be done before any robot revolution can take place. Leitner's team might have won Amazon's challenge last year, but its robotic arm fell to bits on the first day of competition, and most teams experienced technical issues of some kind or another. "These systems are not super robust yet," Leitner says. "If you were to take that system and put it into an Amazon warehouse, I'm not sure how long it would actually work for."

"I'm someone who is exploring a new world," says Brock, "and I'm not ready yet to spring in to the limelight and make industrial applications left and right." But, as Goldberg says, "there's no barrier to putting these into practice. It's already starting to happen." ■

Richard Hodson *is supplements editor for* Nature.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ILLUSTRATION BY SAM FALCONER



Changing landscape

Tax breaks have boosted Australian biotechnology and fuelled overseas investment, but industry insiders fear that threatened budget cuts could slow growth.

BY BIANCA NOGRADY

In 1999, an Australian federal government briefing paper on biotechnology in the country concluded that the sector "hardly rates as an economic force" because of its small size and the financial challenges that it faced in getting products to market.

Now, barely two decades later, Australia has ranked in the top five globally for biotechnology three years running, outperforming nations such as the United Kingdom and Germany, according to *Scientific American Worldview*. The medical-biotechnology sector has benefited from a multibillion-dollar windfall of government funding, as well as from substantial tax breaks for companies investing in research and development.

A run of high-profile international sales and licensing deals has attracted considerable interest from multinational pharmaceutical companies looking to invest in biotechnology. And although the medical sector garners the most headlines and funding, Australian agricultural, environmental and industrial biotechnologies are making their debut on the world stage. But despite the enthusiasm, there are concerns that Australian biotechnology faces a major threat from the federal government's planned overhaul of a tax-incentive scheme that many argue has delivered the industry's biggest boost.

Dean Moss, who leads UniQuest — the commercialization arm of the University of Queensland in Brisbane — highlights the sector's buoyancy, adding: "You've got a very, very healthy environment, you've got a load of state initiatives for medical devices and therapeutics, so there's a real alignment of the planets to make it easier to access capital to develop."

Scientific-research funding has been severely cut in recent years — falling to just below 0.6% of gross domestic product in 2014–15. But biotechnology has received a lot of cash from the government. First, there was the Medical Research Future Fund, announced in May 2014; this was an Aus$20-billion federal-government fund (worth about US$18 billion at the time) to support medical research and innovation. Then, in 2016, came the Biomedical Translation Fund, an $500-million scheme jointly supported by federal government and private coffers to "translate good science and research effort into real-life healthcare solutions".

And the Research and Development (R&D) Tax Incentive has been providing refundable tax offsets since July 2011. Glenn Cross, chief executive of AusBiotech in Melbourne — the sector's industry body — says that not only is the top end of the biotechnology sector looking extremely healthy (see 'Deal makers'), but there's plenty more activity further down. ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

"At the other end of the market, we're continuing to see new technologies commercialized and investment coming into the sector," Cross says. As well as a steady stream of biotechnology companies listing on the Australian stock exchange, Cross says, the R&D Tax Incentive scheme is encouraging international companies to see Australia as a destination for research and development.

Companies with a turnover of less than $20 million a year may be able to get 43.5% of their annual R&D spend refunded in cash; companies with higher turnovers could be eligible for a 38.5% tax offset. In 2013–14, there were nearly 12,000 registrations for the incentive scheme. Overall, the federal government refunded around $2.95 billion for an R&D spend by those companies of $19.5 billion.

But two reports have questioned the scheme's effectiveness in boosting research and development. The first, presented to Australia's prime minister, Malcolm Turnbull, in April 2016, concluded that the growth of the tax incentive was jeopardizing the scheme's long-term sustainability (see go.nature.com/2rj9z1s); it proposed capping a company's annual refund at $2 million, to keep costs down. Its authors argued that this was unlikely to harm small and medium-sized enterprises, which were already gaining a premium refund because of their lower turnover, nor would it significantly affect larger companies, which had greater access to alternative sources of funding.

The second report, published last November, recommended an annual cap of $4 million per company, with a maximum cumulative refund of $40 million per company (see go.nature.com/2rjfgyh). The report, produced by Innovation and Science Australia, an independent advisory board, also suggested that the tax incentive should kick in only once a company's R&D spending exceeds 1% of its total annual expenditure.

The biotechnology sector is unenthusiastic about the capping proposals. "The biggest concern we have is that they would make changes to the R&D Tax Incentive that would stop investment coming in internationally," says Cross. They could also result in less research and fewer clinical trials being conducted in Australia.

Federal Treasurer Scott Morrison warned that the incentive will undergo big changes in the May 2018 budget, saying in an April speech that the scheme "has been taken for a ride by some and integrity needs to be restored."

International interest and investment is vital for the growth of Australian biotechnology. Sue MacLeman, managing director and chief executive of MTPConnect in Melbourne — a not-for-profit, government-funded organization to support the medical-technology and pharmaceutical sectors — says that, because the sector is relatively small on a global scale, Australia is unlikely to develop its own big multinational pharmaceutical or biotechnology companies. Rather, it will fit into the bigger picture, she thinks, as a market of high-quality biotech start-ups for international buyers.

TALENT SPOTTING

"We're seeing an increase in the number of scouts on the ground from those pharma companies, clearly looking for these sorts of [biotechnology] science," MacLeman says. She adds that Australian biotechology companies are increasingly offering technology that has already made it to phase I or phase II clinical trials. This makes a start-up or spin-off much more attractive to international pharmaceutical companies — and that means a bigger return on investment for the Australian company once it is bought, MacLeman says.

> "THERE'S A REAL **ALIGNMENT** OF THE PLANETS TO MAKE IT EASIER TO ACCESS **CAPITAL.**"

Australian medical biotechnology is probably most famous for a cervical-cancer vaccine developed by Ian Frazer and Jian Zhou at the University of Queensland in the 1990s, but the sector is now making waves internationally in a wide variety of fields.

"We're very competitive in oncology, infectious diseases, things like vaccines, respiratory, neurology, ophthalmology [and] we see good progress in regenerative medicine," MacLeman says, adding that Australia's clinical-trial capability is part of its biotechnology success story. A collaborative network of clinical-trial units extends across the country, and states such as Victoria are courting international pharmaceutical companies, urging them to bring their early-stage trials to Australia.

"The fact that we're being asked to do those initial clinical trials here in Australia means that we have the frameworks in place. We have the skill and expertise in place to actually look at some of these technologies," says MacLeman.

"You've got world-class infrastructure and good clinicians, so you're getting good data under independent ethics review and consent, so that the data has integrity and security and you can make meaningful investment decisions about the continued development of those programmes."

MacLeman says that approval for clinical trials through the Australian Clinical Trial Notification scheme can take as little as six weeks, compared with about a year for an application submitted in the United States under the Investigational New Drug programme.

But things are less streamlined for drugs from biological sources, says Melissa Little, programme leader of Stem Cells Australia — a national network of stem-cell researchers and institutions. Little, based in Melbourne, says that "for a clinical trial where you've got to do cellular production, even for a phase 1, it's easier to go offsite". That's because the Therapeutic Goods Administration, Australia's regulatory body for therapeutic products, is "much smaller and much less experienced with biologics" than is the US Food and Drug Administration.

The TGA also has a loophole in its regulation that allows a registered clinician to administer a stem-cell treatment using a patient's own cells — even if that treatment is unproven — without going through any regulatory checks. Organizations such as Stem Cells Australia are lobbying hard to have the issue dealt with, and Little says that the TGA is looking at how to do this.

The regulator is well respected at home and abroad, but it has been criticized for taking too long to approve medicines and medical devices, compared with its overseas counterparts — a concern that was highlighted in a 2015 review commissioned by the government.

Since then, the TGA has set up a programme called SME Assist, to help small and medium-sized enterprises navigate more easily through the registration process.

BRANCHING OUT

Although 69% of Australian biotechnology takes place in the health and biomedical arena, the country is also emerging as a global player in agricultural, environmental and

DEAL MAKERS

Australia's biotechnology sector boasts a series of record-breaking deals in recent years.

MAY 2014
Irish pharma group Shire buys Fibrotech and its innovative fibrosis therapeutic candidates for Aus$615.1 million (US$557.5 million).

JUNE 2015
Novartis buys Spinifex Pharmaceuticals for an estimated $1 billion — the biggest Australian biotechnology deal up to that time.

SEPT 2015
Dr Reddy's, founded in India in 1984, acquires Hatchtech and its head-lice treatments for $279 million.

Nanotech drug-delivery company Starpharma signs a licensing deal with AstraZeneca worth more than $650 million.

AUGUST 2016
Protagonist Therapeutics, a University of Queensland spin-off, raises $118 million in Nasdaq listing.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

industrial biotechnology. Michael Christie, a patent attorney in Sydney who chairs the New South Wales division of the Ag Institute Australia — the industry body for agricultural and natural-resource management professionals — says that the country is particularly strong on basic and applied research in agricultural biotechnology. But he adds that the take-up is more patchy. For one thing, South Australia, Tasmania and the Australian Capital Territory all have moratoriums on the commercial cultivation of genetically modified crops.

Christie says that the regulatory framework for genetically modified crops is also a barrier to adoption. "Generally speaking, the scheme is quite complicated and the procedures for obtaining regulatory approval are unduly cumbersome," he says. But he also stresses that the framework helps to maintain Australia's reputation as a "clean and green" source of agricultural products.

Although it has not received the same funding and investment as medical biotechnology, Christie says that agricultural biotechnology is becoming more of a topic of conversation among investors. Much of the research in the field is done within the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia's national science agency. But in recent years, the formation of primary industry-focused research and development corporations, such as AgriFutures Australia, has widened the field.

"These groups have placed a stronger emphasis on commercial outcomes, so we're seeing more applied-research programs as well as collaborations being formed between publicly funded research organizations and industry," Christie says.

CSIRO is also a hotbed of environmental and industrial biotechnology — particularly in relation to mining, an area in which Australia is a key global player — says Anna Kaksonen, who leads the biotechnology and synthetic-biology group at CSIRO in Perth.

Mining biotechnology uses microbes "to extract metals from low-grade ores and also metal-containing waste material such as e-waste, battery waste, mining waste," Kaksonen explains. It is also being used to clean up polluted sites, such as old mines. "There are a lot of legacy sites in Australia, so there is a lot of work in that space where microbes could be used to remove sulfate metals and acidity from the water," she says.

Many of the microbes used in this work are found in nature, but researchers are looking at ways of engineering them to better withstand the their contaminated environments.

SKILLS GAP

Australia has no shortage of ideas in biotechnology, but it is struggling to find people with relevant commercial experience, says Cross, particularly at the most senior leadership levels.

"We're certainly getting a lot more of our senior executives who have gone overseas and worked in big pharma companies or big medtech companies coming back to Australia," he says. "But the ability to satisfy all of the senior executive-level jobs that are coming up in the industry — it's difficult to do just from Australia."

To address this, the Australian government announced in March this year a pilot of a new version of the entrepreneur visa. Unlike the current visa, this one does not require entrepreneurs to have capital backing. Ministers also launched a Global Talent Scheme, offering four-year Temporary Skill Shortage visas to science, engineering or medicine start-ups, as well as to established businesses seeking people for positions with salaries above $180,000.

The relatively small size of Australia's private biotechnology sector — the national Stock Exchange lists 140 life-sciences companies — can also make it difficult for graduates to find work outside academia and research institutions. Andrew Webb, acting division head for systems biology and personalized medicine at the Walter and Eliza Hall Institute of Medical Research in Melbourne, says that even positions in academic research are "few and far between" in Australia.

"I don't know where postdocs are going, but there's certainly not a huge amount of jobs available, certainly in the academic sector."

Australia's commercial biotechnology sector is growing fast, however, and hotspots such as Melbourne — which hosts 27 leading biotechnology-research institutes — are gaining ground internationally. In terms of supporting and nurturing biotechnology start-ups and spin-offs, Webb says that Melbourne is approaching the standard of Boston, San Francisco and Cambridge, Massachusetts.

Webb says that, despite Australia's difficulties in commercializing its research, the speed of development makes biotechnology an exciting space to be in. "Everything is advancing so rapidly that there are enormous opportunities in this space." ■

NOV 2016
SUDA, a company that specializes in oro-mucosal drug delivery, signs deal worth up to $34 million with Chinese pharmaceutical company Eddingpharm for an insomnia therapy.

JULY 2017
The Walter and Eliza Hall Institute in Melbourne sells a royalty stake in its anti-cancer treatment venetoclax to a subsidiary of the Canada Pension Plan Investment Board for up to $424 million.

FEBRUARY 2018
Pharma group MSD pays around $502 million to acquire Viralytics — a spin-off from the University of Newcastle focused on the cancer-busting coxsackievirus A21.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.